

# Sparse Tensor Response Regression and Neuroimaging Analysis

Will Wei Sun and Lexin Li

University of Miami and University of California at Berkeley

## Abstract

Motivated by applications in neuroimaging analysis, we propose a new regression model with a tensor response and a vector predictor. The model embeds two key sparse structures: element-wise sparsity and low-rankness. It can handle both a general and a symmetric tensor response, and thus is applicable to both structural and functional neuroimaging data. We formulate the model parameter estimation as a non-convex optimization problem, and develop an efficient alternating updating algorithm. We establish a non-asymptotic estimation error bound for the actual estimator obtained from the proposed algorithm. This error bound reveals an interesting interaction between the computational efficiency and the statistical rate of convergence. Based on this general error bound, we further obtain an optimal estimation error rate when the distribution of the error tensor is Gaussian. We illustrate the efficacy of our model through intensive simulations and an analysis of the Autism spectrum disorder neuroimaging data.

**Key Words:** Functional connectivity analysis, Magnetic resonance imaging, Non-asymptotic error bound, Non-convex optimization, Tensor decomposition.

---

<sup>1</sup>Will Wei Sun is Assistant Professor, Department of Management Science, University of Miami School of Business Administration, Miami, FL 33146. Email: wsun@bus.miami.edu. Lexin Li is Associate Professor, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-3370. Email: lexinli@berkeley.edu.

# 1 Introduction

In this article, we study the regression model with a tensor response and a vector predictor. Tensor, a.k.a. multidimensional array, is now frequently arising in a wide range of scientific and business applications ([Karatzoglou et al., 2010](#); [Zheng et al., 2010](#); [Liu et al., 2013](#); [Zhou et al., 2013](#); [Signoretto et al., 2014](#); [He et al., 2014](#), among many others). Our motivation comes from neuroimaging analysis. One example is anatomical magnetic resonance imaging (MRI), where the data takes the form of a three-dimensional array, and image voxels correspond to brain spatial locations. It is of keen interest to compare the MRI scans of brains between the subjects with neurological disorder and the healthy controls after adjusting for additional covariates such as age and gender. Another example is functional magnetic resonance imaging (fMRI). Scientifically it is crucial to compare brain connectivity patterns across different diagnostic groups, where the connectivity is encoded by a symmetric matrix, with rows and columns corresponding to brain regions, and entries corresponding to interaction and dependency between those brain regions. Both can be more generally formulated as a regression problem, with image tensor or connectivity matrix serving as a response, and the group indicator and other covariates forming a predictor vector.

## 1.1 Our proposal

We develop a general class of tensor response regression models and embed two key sparse structures: element-wise sparsity and low-rankness. Both structures serve to greatly reduce the computational complexity of the estimation procedure. Meanwhile, both are scientifically plausible in the neuroimaging and plenty of other applications, and have been widely employed in high-dimensional tensor regressions (e.g., [Zhou et al., 2013](#); [Zhu et al., 2014](#); [Raskutti and Yuan, 2016](#)). On the other hand, a unique feature of our proposal is that, it can not only handle a general tensor response, but also a symmetric tensor response, and thus is applicable to both structural and functional neuroimaging analysis.

We formulate the learning of our tensor response model as a non-convex optimization problem, and accordingly develop an efficient alternating updating algorithm. Our algorithm consists of two major steps, and each step iteratively updates a subset of the unknown

parameters while fixing the others. In Step 1, we reformulate the estimation procedure as a sparse tensor decomposition problem and then employ a recently proposed truncated tensor power method (Anandkumar et al., 2014; Sun et al., 2016) for fast updating. In Step 2, we utilize the bi-convex structure of the problem to obtain a closed-form solution for the update of the rest of the parameters.

We carry out a non-asymptotic theoretical analysis for the actual estimator obtained from our optimization algorithm. Based upon a set of our newly developed techniques to tackle the non-convexity in estimation, we obtain an explicit error bound of the estimator. Specifically, let  $\mathcal{E}_i, i = 1, \dots, n$ , denote the error tensor, and  $\Theta^*$  denote the set of all true parameters. Given an initial parameter with an initialization error  $\epsilon$ , the finite sample error bound of the  $t$ -th step solution  $\hat{\Theta}^{(t)}$  consists of two parts:

$$D\left(\hat{\Theta}^{(t)}, \Theta^*\right) \leq \underbrace{\kappa^t \epsilon}_{\text{computational error}} + \underbrace{\frac{1}{1-\kappa} \max \left\{ C \cdot \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right), \frac{\tilde{C}}{\sqrt{n}} \right\}}_{\text{statistical error}},$$

with a high probability. Here  $\kappa \in (0, 1)$  is a contraction coefficient,  $C$  and  $\tilde{C}$  are some positive constants, and  $\eta(n^{-1} \sum_{i=1}^n \mathcal{E}_i, s)$  represents the  $s$ -sparse spectral norm of the averaged error tensor; see (8) for a formal definition of this norm. This error bound portrays the estimation error in each iteration, and reveals an interesting interplay between the computational efficiency and the statistical rate of convergence. Note that the computational error decays geometrically with the iteration number  $t$ , whereas the statistical error remains the same when  $t$  grows. Therefore, this bound provides a meaningful choice of the maximal number of iterations  $T$ , such that the computational error is dominated by the statistical error, and the error bound is in the same order of the statistical error. After  $T$  iteration steps, the estimator from our algorithm falls within the statistical precision of the true parameter  $\Theta^*$ .

Additionally, this finite sample error bound also provides a general theoretical guarantee of our estimator, and the result holds for any distribution of the error tensor  $\mathcal{E}_i$  by noting that it relies on  $\mathcal{E}_i$  only through its sparse spectral norm  $\eta(n^{-1} \sum_{i=1}^n \mathcal{E}_i, s)$ . In Sections 4.3 and 4.4, we obtain explicit forms of statistical errors when the distribution of  $\mathcal{E}_i$  is available. In particular, when the third-order error tensor  $\mathcal{E}_i \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ ,  $i = 1, \dots, n$ , follows an i.i.d.

Gaussian distribution, we have

$$D\left(\widehat{\Theta}^{(T)}, \Theta^*\right) = O_p\left(\sqrt{\frac{s^3 \log(d_1 d_2 d_3)}{n}}\right),$$

where  $s$  is the cardinality parameter of the decomposed components in the tensor coefficients. This rate is optimal, by noting that there are at most  $O(s^3)$  non-zero elements and  $O(d_1 d_2 d_3)$  free parameters in the tensor coefficient. When the mode of the tensor is one, the tensor response regression reduces to the  $d$ -dimensional vector regression and our statistical error reduces to  $O_p(\sqrt{s \log d/n})$ , which is known to be minimax optimal (Wang et al., 2014b; Yi and Caramanis, 2015).

## 1.2 Related works and our contributions

Our work is related to but also clearly distinctive from a number of recent developments in tensor decomposition and tensor regression.

The first is a class of tensor decomposition methods (Chi and Kolda, 2012; Liu et al., 2012; Anandkumar et al., 2014; Yuan and Zhang, 2016; Sun et al., 2016). Tensor decomposition is an unsupervised learning method that aims to find the best low-rank approximation of a single tensor. Our proposed tensor response regression, however, is a supervised learning method, which aims to estimate the coefficient tensor that characterizes the association between the tensor response and the vector predictor. Although we utilize the sparse tensor decomposition step of Sun et al. (2016) in our estimation algorithm, our objectives and technical tools involved are completely different from Sun et al. (2016). Particularly, because we work with multiple tensor samples, the consistency of our estimator is indexed by both the tensor dimension and the sample size. This is different from the tensor decomposition estimator in Sun et al. (2016) that works with a single tensor only. In addition, our algorithm can be coupled with other sparse tensor decomposition algorithms as well.

The second related line of research tackles tensor regression where the response is a scalar and the predictor is a tensor (Zhou et al., 2013; Wang et al., 2014a; Wang and Zhu, 2016). In particular, Zhou et al. (2013) imposed a low-rank structure on the coefficient tensor, similar to the low-rank principle we also employ. However, Zhou et al. (2013) treated the

tensor as a predictor, whereas we treat it as a response. The resulting regression models thus have different focus and interpretations. The tensor predictor regression focuses on understanding the change of a clinical outcome as the tensor image varies, while the tensor response regression aims to study the change of the image as the predictors such as the disease status and age vary. The two also involve utterly different techniques when it comes to theoretical analysis. In a way, their difference is in analogous to that between multi-response regression and multi-predictor regression in the classical vector-valued regression context.

The third and more relevant line of works directly studies regression with a tensor response (Zhu et al., 2009; Li et al., 2011; Li and Zhang, 2016; Raskutti and Yuan, 2016). In particular, Zhu et al. (2009) considered a  $3 \times 3$  symmetric positive definite matrix arising from diffusion tensor imaging as a response, and developed an intrinsic regression model by mapping the Euclidean space of covariates to the Riemannian manifold of positive-definite matrices. Unlike Zhu et al. (2009), we consider a general or symmetric tensor response in the Euclidean space, and allow the dimension of the tensor response to diverge with the sample size. Li et al. (2011) estimated regression parameters by building iteratively increasing neighbors around each voxel and smoothing observations within the neighbors with weights. By contrast, we model all the voxels in an image tensor in a joint fashion.

Li and Zhang (2016) recently proposed an envelope-based tensor response model, and our work differs from theirs in several ways. First, Li and Zhang (2016) utilized a generalized sparsity principle to exploit the redundant information in the tensor response, by seeking linear combinations of the response that are irrelevant to the regression. Our method instead utilizes the classical sparsity in terms of the individual entries. As such we can directly identify brain subregions that are affected by the disorder, and is easier to interpret, whereas Li and Zhang (2016) can not. Second, Li and Zhang (2016) obtained the  $\sqrt{n}$ -convergence rate for the *global* minimizer of their objective function when the tensor dimension is fixed. However, their objective function is non-convex and there is no guarantee that the optimization algorithm can find this global minimizer. By contrast, we derive the error rate of the *actual* minimizer of our algorithm at each iteration, and we also permit a diverging tensor dimension. Third, Li and Zhang (2016) could not directly incorporate the symmetry constraint when the tensor

response is symmetric, which is often encountered in functional imaging analysis. To the best of our knowledge, our solution is the first that can simultaneously tackle both a general and a symmetric tensor response in a regression setup.

[Raskutti and Yuan \(2016\)](#) developed a class of regression models, under the assumptions of low-dimensionality and Gaussian errors, when either or both the response and predictor are tensors. Different from our solutions, [Raskutti and Yuan \(2016\)](#) formulated the problem as a convex optimization under a crucial assumption that the regularizer was convex and weakly decomposable. We do not require this assumption, and instead tackle a non-convex optimization problem. Consequently, a new set of proof techniques is required for our theoretical analysis. Moreover, in [Raskutti and Yuan \(2016\)](#), the low-rankness of the estimator is achieved via a tensor nuclear norm. In spite of its convexity, tensor nuclear norm is known to be computationally NP-hard ([Friedland and Lim, 2014](#)), which is likely to impede its practical application.

Finally, we develop a set of tools for the theoretical analysis of non-convex optimization, which is notably different from recent developments in this area ([Wang et al., 2014b](#); [Yi and Caramanis, 2015](#); [Sun et al., 2015](#); [Balakrishnan et al., 2016](#)). A common technique used in those non-convex optimization analysis is to separately establish the convergence for the population and sample optimizers then combine the two. By contrast, our analysis hinges on exploitation of the bi-convex structure of the objective function, as well as a careful characterization of the impact of the intermediate sparse tensor decomposition step on the estimation error in each iteration step. The bi-convex structure frequently arises in many optimization problems. As such the tools we develop are also of independent interest, and enrich and expand current theoretical analysis of non-convex optimization.

### 1.3 Notations and structure

We adopt the following notations throughout this paper. Denote  $[d] = \{1, \dots, d\}$ , and  $\mathbf{I}_d$  the  $d \times d$  identity matrix. Let  $1(\cdot)$  denote the indicator function, and  $\circ$  the outer product between vectors. For a vector  $\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_d)^\top \in \mathbb{R}^d$ ,  $\|\mathbf{v}\|$  refers to its Euclidean norm, and  $\|\mathbf{v}\|_0$  denotes the number of nonzero entries in  $\mathbf{v}$ . For a matrix  $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{d \times d}$ ,  $\|\mathbf{A}\|$  denotes its

spectral norm. We say  $a_n = o(b_n)$  if  $a_n/b_n$  converges to 0 as  $n$  diverges. We use  $C_0, C_1, \dots$  to denote generic absolute constants, while the values of these constants may vary.

Following [Kolda and Bader \(2009\)](#), we adopt the following tensor operations. For a tensor  $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ , we define  $\mathcal{T}_{i,j,l}$  as its  $(i, j, l)$ -th entry. For vectors  $\mathbf{u} \in \mathbb{R}^{d_1}, \mathbf{v} \in \mathbb{R}^{d_2}, \mathbf{w} \in \mathbb{R}^{d_3}$ , we define the multilinear combination of the tensor entries as  $\mathcal{T} \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} := \sum_{i \in [d_1]} \sum_{j \in [d_2]} \sum_{l \in [d_3]} \mathbf{u}_i \mathbf{v}_j \mathbf{w}_l \mathcal{T}_{i,j,l} \in \mathbb{R}$ . We define the tensor spectral norm as  $\|\mathcal{T}\| := \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1} |\mathcal{T} \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w}|$ , and tensor Frobenius norm as  $\|\mathcal{T}\|_F := \sqrt{\sum_{i,j,l} \mathcal{T}_{i,j,l}^2}$ . For two tensors  $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ , we define their inner product as  $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,l} \mathcal{A}_{i,j,l} \mathcal{B}_{i,j,l}$ .

The rest of the article is organized as follows. Section 2 introduces the proposed sparse low-rank tensor response regression model, and Section 3 presents the optimization algorithm. Section 4 establishes the estimation error bound. Section 5 presents the simulation results, and Section 6 applies our method to an analysis of the Autism spectrum disorder imaging data. The appendix and the online supplement collect all technical proofs.

## 2 Model

For an  $m$ th-order tensor response  $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times \dots \times d_m}$ , and a vector of predictors  $\mathbf{x}_i \in \mathbb{R}^p$ ,  $i = 1, \dots, n$ , we consider the tensor response regression model of the form,

$$\mathcal{Y}_i = \mathcal{B}^* \times_{m+1} \mathbf{x}_i + \mathcal{E}_i, \quad (1)$$

where  $\mathcal{B}^* \in \mathbb{R}^{d_1 \times \dots \times d_m \times p}$  is an  $(m+1)$ th-order tensor coefficient, and  $\mathcal{E}_i \in \mathbb{R}^{d_1 \times \dots \times d_m}$  is an error tensor independent of  $\mathbf{x}_i$ . Throughout the paper, we also use  $d_{m+1} := p$  to represent the dimension of the predictor vector  $\mathbf{x}_i$ . Our goal is to estimate  $\mathcal{B}^*$  given  $n$  i.i.d. observations  $\{(\mathbf{x}_i, \mathcal{Y}_i), i = 1, \dots, n\}$ .

To facilitate estimation of the ultrahigh dimensional unknown parameters under a limited sample size, it is crucial to introduce some sparse structures. Two most commonly used structures are element-wise sparsity and low-rankness ([Raskutti and Yuan, 2016](#)). In the context of tensor response regression, we assume that  $\mathcal{B}^*$  admits the following rank- $K$  decomposition structure ([Kolda and Bader, 2009](#)),

$$\mathcal{B}^* = \sum_{k \in [K]} w_k^* \beta_{k,1}^* \circ \dots \circ \beta_{k,m}^* \circ \beta_{k,m+1}^*, \quad w_k^* \in \mathbb{R}, \beta_{k,j}^* \in \mathbb{S}^{d_j}, \quad (2)$$

where  $\mathbb{S}^d = \{\mathbf{v} \in \mathbb{R}^d \mid \|\mathbf{v}\| = 1\}$ , and  $\|\boldsymbol{\beta}_{k,j}^*\|_0 \leq d_{j0} < d_j$  for any  $k \in [K]$  and  $j \in [m+1]$ . Moreover, we assume  $w_{\max}^* = w_1^* \geq \dots \geq w_K^* = w_{\min}^* > 0$  and assume each  $w_i^*$  to be bounded away from 0 and  $\infty$ . Under the structure of (2), estimating  $\mathcal{B}^*$  reduces to the estimation of  $w_k^*, \beta_{k,1}^*, \dots, \beta_{k,m+1}^*$  for any  $k \in [K]$ , and the corresponding parameter space becomes  $\mathbb{B} := \{w_k \in \mathbb{R}, \boldsymbol{\beta}_{k,j} \in \mathbb{R}^{d_j}, k \in [K], j \in [m+1] \mid c_1 \leq |w_k| \leq c_2, \|\boldsymbol{\beta}_{k,j}\|_2 = 1, \|\boldsymbol{\beta}_{k,j}\|_0 \leq d_{j0}\}$ . Accordingly, the number of unknown parameters is reduced from  $p \prod_{j=1}^m d_j$  to  $K(\sum_{j=1}^m d_j + p)$ .

Furthermore, we assume the coefficient tensor  $\mathcal{B}^*$  is sparse in that the individual entries of  $\mathcal{B}^*$  are zero. This sparsity assumption would further reduce the number of unknown parameters, and also facilitate the interpretation of the final estimator. Toward that end, we propose to solve the following constrained optimization problem,

$$\begin{aligned} \min_{\substack{w_k, \beta_{k,1}, \dots, \beta_{k,m+1} \\ k \in [K], j \in [m+1]}} \quad & \frac{1}{n} \sum_{i=1}^n \left\| \mathcal{Y}_i - \sum_{k \in [K]} w_k (\beta_{k,m+1}^\top \mathbf{x}_i) \beta_{k,1} \circ \dots \circ \beta_{k,m} \right\|_F^2, \\ & \text{subject to } \|\beta_{k,j}\|_2 = 1, \|\beta_{k,j}\|_0 \leq s_j, \end{aligned} \quad (3)$$

where  $s_j$  is the cardinality parameter of the  $j$ -th component. Here we encourage the sparsity of the decomposed components via a hard-thresholding penalty to control  $\|\beta_{k,j}\|_0$ . Compared to the lasso type penalized approach in sparse learning, the hard-thresholding method avoids bias and has been shown to be more appealing in numerous high-dimensional learning problems (Wang et al., 2014b; Sun et al., 2016).

### 3 Estimation

The problem in (3) is a non-convex optimization. The key of our estimation procedure is an exploration of its bi-convex structure. In particular, the objective function in (3) is bi-convex in  $(\beta_{k,1}, \dots, \beta_{k,m+1})$ ,  $k \in [K]$ , in that it is convex in  $\beta_{k,j}$  when all other parameters are fixed. Utilizing this property, we propose an efficient alternating updating algorithm to solve (3).

#### 3.1 Algorithm

We first summarize our estimation procedure in Algorithm 1, then present the details of its two key steps.



---

**Algorithm 1** Alternating updating algorithm for sparse low-rank tensor response regression estimation.

---

- 1: **Input:** data  $\{(\mathbf{x}_i, \mathcal{Y}_i), i = 1, \dots, n\}$ , rank  $K$ , cardinality vector  $(s_1, \dots, s_m)$ .
  - 2: **Initialize**  $w_k = 1$  and random unit-norm vectors  $\beta_{k,1}, \dots, \beta_{k,m+1}$  for each  $k \in [K]$ .
  - 3: **Until** termination condition holds, **Do**
  - 4:   Step 1: For  $k = 1$  to  $K$ , compute  $\mathcal{R}_i$  in (5) and solve  $\hat{w}_k, \hat{\beta}_{k,1}, \dots, \hat{\beta}_{k,m}$  in (6) via the sparse tensor decomposition (Sun et al., 2016) with parameters  $K$  and  $(s_1, \dots, s_m)$ .
  - 5:   Step 2: For  $k = 1$  to  $K$ , update  $\hat{\beta}_{k,m+1}$  according to Lemma 2.
  - 6: **Output:**  $\hat{w}_k, \hat{\beta}_{k,1}, \dots, \hat{\beta}_{k,m+1}$  for each  $k \in [K]$ .
- 

**Step 1:** The first core step of our algorithm is to update  $w_k, \beta_{k,1}, \dots, \beta_{k,m}$  for each  $k = 1, \dots, K$ , given  $\beta_{j,m+1}, j = 1, \dots, K$  and  $w_{k'}, \beta_{k',1}, \dots, \beta_{k',m}, k' \neq k$ . Letting  $\alpha_{ik} := \beta_{k,m+1}^\top \mathbf{x}_i$ ,  $i = 1, \dots, n$ , we note that solving (3) is equivalent to

$$\min_{\substack{w_k, \beta_{k,1}, \dots, \beta_{k,m} \\ \|\beta_{k,j}\|_2=1, \|\beta_{k,j}\|_0 \leq s_j, j \in [m]}} \frac{1}{n} \sum_{i=1}^n \alpha_{ik}^2 \left\| \mathcal{R}_i - w_k \beta_{k,1} \circ \dots \circ \beta_{k,m} \right\|_F^2, \quad (4)$$

where the residual tensor term  $\mathcal{R}_i$  is of the form,

$$\mathcal{R}_i := \left[ \mathcal{Y}_i - \sum_{k' \neq k, k' \in [K]} w_{k'} \alpha_{ik'} \beta_{k',1} \circ \dots \circ \beta_{k',m} \right] / \alpha_{ik}. \quad (5)$$

The next lemma shows that one can solve (4) via an efficient sparse tensor decomposition procedure as developed in Sun et al. (2016).

**Lemma 1.** Denote  $\bar{\mathcal{R}} := \frac{1}{n} \sum_{i=1}^n \alpha_{ik}^2 \mathcal{R}_i$ . The solution of (4) is also the solution of

$$\min_{\substack{w_k, \beta_{k,1}, \dots, \beta_{k,m} \\ \|\beta_{k,j}\|_2=1, \|\beta_{k,j}\|_0 \leq s_j, j \in [m]}} \left\| \bar{\mathcal{R}} - w_k \beta_{k,1} \circ \dots \circ \beta_{k,m} \right\|_F^2. \quad (6)$$

Lemma 1 implies that the optimization problem in (4) reduces to a sparse rank-one tensor decomposition on the averaged tensor  $\bar{\mathcal{R}}$ . To efficiently solve (6), in this paper we employ a truncation-based sparse tensor decomposition (Sun et al., 2016), by first solving the non-sparse tensor decomposition components and then truncating them to achieve desirable sparsity. It is also noteworthy that our method is flexible in the choice of optimization algorithm for solving (6), and Algorithm 1 can be coupled with other sparse tensor decomposition algorithms, e.g., Chi and Kolda (2012); Liu et al. (2012).

When the tensor response  $\mathcal{Y}_i$  is symmetric, the resulting coefficient  $\mathcal{B}$  should be symmetric too. Our algorithm can easily adapt to this scenario, by setting  $\beta_{k,1} = \dots = \beta_{k,m} = \beta_k$  for each  $k \in [K]$ , and the cardinality parameters  $s_1 = \dots = s_m = s$ . That is, we slightly modify Step 1 in Algorithm 1 as: for  $k = 1$  to  $K$ , compute  $\mathcal{R}_i$  in (5) and solve  $\hat{w}_k, \hat{\beta}_k$  via the symmetric sparse tensor decomposition (Sun et al., 2016) with parameters  $K$  and  $s$ .

**Step 2:** The second core step of our algorithm is to update  $\beta_{k,m+1}$  for each  $k = 1, \dots, K$ , given  $w_j, \beta_{j,1}, \dots, \beta_{j,m}$ ,  $j = 1, \dots, K$  and  $\beta_{k',m+1}$ ,  $k' \neq k$ . Letting  $\mathcal{A}_k = w_k \beta_{k,1} \circ \dots \circ \beta_{k,m}$ , solving (3) is then equivalent to find

$$\hat{\beta}_{k,m+1} := \arg \min_{\alpha} \frac{1}{n} \sum_{i=1}^n \left\| \mathcal{T}_i - \alpha^\top \mathbf{x}_i \mathcal{A}_k \right\|_F^2, \quad (7)$$

where the residual tensor  $\mathcal{T}_i = \mathcal{Y}_i - \sum_{k' \neq k, k' \in [K]} w_{k'} (\beta_{k',m+1}^\top \mathbf{x}_i) \beta_{k',1} \circ \dots \circ \beta_{k',m}$ . In our context of neuroimaging applications, the dimension of the predictor vector is usually small; e.g.,  $p = 3$  in the real data analysis in Section 6. Therefore we have chosen not to introduce any additional sparsity constraint on  $\beta_{k,m+1}$ , though it is straightforward to incorporate such constraint if one chooses to do so. Then the next lemma gives a closed-form solution for the update of  $\hat{\beta}_{k,m+1}$ .

**Lemma 2.** *The solution of (7) is given by*

$$\hat{\beta}_{k,m+1} = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \frac{n^{-1} \sum_{i=1}^n \langle \mathcal{T}_i, \mathcal{A}_k \rangle \mathbf{x}_i}{\|\mathcal{A}_k\|_F^2}.$$

In this closed-form solution,  $n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$  corresponds to the sample covariance matrix. When the sample size  $n$  is larger than the dimension  $p$  of the predictor vector, this term is invertible. This is often the case in the neuroimaging applications, where  $n$  is usually in tens or hundreds while  $p$  is three to five. When this inverse does not exist, one may employ a sparse graphical model approach (Yuan and Lin, 2007; Friedman et al., 2008) and replace this inverse with a sparse estimate of the precision matrix.

Finally, we terminate the alternating update of Steps 1 and 2 when the new estimates are close to the ones from the previous iteration. The termination condition is set as

$$\max_{j \in [m+1], k \in [K]} \min \left\{ \|\hat{\beta}_{k,j}^{(t)} - \hat{\beta}_{k,j}^{(t-1)}\|, \|\hat{\beta}_{k,j}^{(t)} + \hat{\beta}_{k,j}^{(t-1)}\| \right\} \leq 10^{-4}.$$

### 3.2 Tuning parameter selection

In Algorithm 1, the rank  $K$  and the cardinality  $s_1, \dots, s_m$  are tuning parameters. We propose to select those parameters via a BIC-type criterion. Specifically, given a pre-specified set of rank values  $\mathcal{K}$  and a pre-specified set of cardinality values  $\mathcal{S}_1, \dots, \mathcal{S}_m$ , we choose the combination of parameters  $(\hat{K}, \hat{s}_1, \dots, \hat{s}_m)$  that minimizes

$$\text{BIC} = \log \left( \sum_{i=1}^n \|\mathcal{Y}_i - \hat{\mathcal{B}} \times_{m+1} \mathbf{x}_i\|_F^2 \right) + \frac{\log(n \prod_{j=1}^m d_m)}{n \prod_{j=1}^m d_m} \sum_{k=1}^K \sum_{j=1}^m \|\hat{\beta}_{k,j}\|_0.$$

This criterion balances between model fitting and model sparsity, and a similar version has been commonly employed in the tuning procedure of tensor decompositions (Zhou et al., 2013; Sun et al., 2016).

## 4 Theory

In this section, we establish the error bound of the actual estimator obtained from our Algorithm 1 for the proposed sparse low-rank tensor response regression. The established error bound consists of two quantities: a computational error and a statistical error. The computational error captures the error caused by the non-convexity of the optimization problem, whereas the statistical error measures the error due to finite samples.

In order to compute the distance between the estimator and the truth, we define the distance measure between two unit vectors  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$  as

$$D(\mathbf{u}, \mathbf{v}) := \sqrt{1 - (\mathbf{u}^\top \mathbf{v})^2}.$$

We then have  $D(\mathbf{u}, \mathbf{v}) \leq \min\{\|\mathbf{u} - \mathbf{v}\|, \|\mathbf{u} + \mathbf{v}\|\} \leq \sqrt{2}D(\mathbf{u}, \mathbf{v})$ . The distance function  $D(\mathbf{u}, \mathbf{v})$  resolves the sign issue in the decomposition components since changing the signs of any two component vectors while fixing other component vectors does not affect the generated tensor.

### 4.1 Assumptions

We first introduce the technical assumptions required to guarantee the desirable error bound. The first assumption is about the structure of the true model.

**Assumption 1** (Model Assumption). *In model (1), we assume the true coefficient  $\mathcal{B}^*$  is sparse and low-rank satisfying (2) and such decomposition is unique up to a permutation. Assume  $\|\mathcal{B}^*\| \leq C_1 w_{\max}^*$ , and for each  $i$ , assume  $\|\mathbf{x}_i\| \leq C_2$  and  $|\beta_{k,m+1}^{*\top} \mathbf{x}_i| \geq C_3$  almost surely, for some positive constants  $C_1, C_2, C_3$ . Furthermore, we require the eigenvalues of the sample covariance  $\Sigma := n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$  to satisfy  $c_0 < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < \tilde{c}_0$  for some constants  $c_0, \tilde{c}_0$ .*

The above unique decomposition condition is to ensure the identifiability of tensor decomposition. Kruskal (1976, 1977) provided the classical conditions for such identifiability if the sum of the Kruskal ranks of the  $m$  decomposed component matrices is larger than  $2K+2$ . In our model, the rank  $K$  is fixed, and hence the identifiability of our tensor decomposition (2) is guaranteed when the decomposed components are not highly dependent. Moreover, the conditions on the predictor vector  $\mathbf{x}_i$  are mild and trivially hold when the dimension of  $\mathbf{x}_i$  is fixed. For instance, in Section 6, the dimension of  $\mathbf{x}_i$  is 3 in our neuroimaging data example.

The second assumption is about the initialization in Algorithm 1.

**Assumption 2** (Initialization Assumption). *Define the initialization error*

$$\epsilon := \max \left\{ \max_k \|\hat{w}_k^{(0)} - w_k^*\|_2, \max_{k,j} D(\hat{\beta}_{k,j}^{(0)}, \beta_{k,j}^*) \right\}.$$

*We assume that*

$$\epsilon < \min \left\{ \sqrt{\frac{w_{\min}^*}{2(w_{\min}^* + w_{\max}^* C_1)}}, \frac{w_{\min}^{*2}}{8\sqrt{5}w_{\max}^* C_1 (6\sqrt{2}w_{\min}^* + 2)}, \frac{C_3}{2C_2}, \frac{w_{\min}^*}{2} \right\},$$

*where the constants  $C_1, C_2, C_3$  are as defined in Assumption 1.*

Note that our assumption on the initialization parameters only requires the error to be bounded by some constant; i.e., the initial estimators are not too far away from the true parameters. This assumption is necessary to handle the non-convexity of the optimization, and has been commonly imposed in non-convex optimization (Wang et al., 2014b; Sun et al., 2016; Balakrishnan et al., 2016). It is also noteworthy that this assumption is satisfied by the estimators from sparse singular value decomposition of the unfolding matrix from the original tensor in Step 1 of Algorithm 1; see Sun et al. (2016) for more discussion on the theoretical guarantees of such an initialization procedure.

The third assumption requires that the error tensor  $\mathcal{E}_i$  is controlled. Before stating the assumption, we introduce a critical concept to measure the noise level in the tensor response regression. In particular, we define the sparse spectral norm of  $\mathcal{E}$  as

$$\eta(\mathcal{E}, d_{01}, \dots, d_{0m}) := \sup_{\substack{\|\mathbf{u}_1\|=\dots=\|\mathbf{u}_m\|=1 \\ \|\mathbf{u}_1\|_0 \leq d_{01}, \dots, \|\mathbf{u}_m\|_0 \leq d_{0m}}} \left| \mathcal{E} \times_1 \mathbf{u}_1 \times_2 \dots \times_m \mathbf{u}_m \right|. \quad (8)$$

This quantifies the perturbation error in a sparse scenario, and in the sparse case with  $d_{0j} \ll d_j$  ( $j = 1, \dots, m$ ), it is much smaller than the spectral norm  $\|\mathcal{E}\|$ , i.e.,  $\eta(\mathcal{E}, d_1, \dots, d_m)$ . Denoting  $d_0 = \max_j d_{0j}$ , we have  $\eta(\mathcal{E}, d_{01}, \dots, d_{0m}) \leq \eta(\mathcal{E}, d_0, \dots, d_0)$ . We write  $\eta(\mathcal{E}, d_0) := \eta(\mathcal{E}, d_0, \dots, d_0)$ .

**Assumption 3** (Error Assumption). *Let  $s := \max\{s_1, \dots, s_m\}$ , with  $s_j$  is the cardinality parameter in Algorithm 1. We assume the error tensor  $\mathcal{E}_i$ ,  $i = 1, \dots, n$ , satisfies that*

$$\eta\left(\frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s\right) \leq \frac{C_3 w_{\min}^*}{8},$$

where  $C_3$  is as defined in Assumption 1.

This assumption requires that the observed tensor responses,  $\mathcal{Y}_i$ ,  $i = 1, \dots, n$ , are not too noisy. It is a minor condition, since we only require the spectral norm of the error tensor to be bounded. As we later show in Section 4.3, when  $\mathcal{E}_i$  is a standard Gaussian tensor, its spectral norm converges to zero when the sample size  $n$  increases.

## 4.2 Main result

Next we present our main theorem that describes the estimation error of the actual estimator at each iteration from Algorithm 1.

**Theorem 1.** *Assuming Assumptions 1, 2, and 3, the estimator in the  $t$ -th iteration of Algorithm 1, with  $s \geq d_0 := \max_j d_{j0}$ , satisfies that*

$$\begin{aligned} & \max \left\{ \max_k \|\hat{w}_k^{(t)} - w_k^*\|_2, \max_{k,j} D\left(\hat{\beta}_{k,j}^{(t)}, \beta_{k,j}^*\right) \right\} \\ & \leq \underbrace{\kappa^t \epsilon}_{\text{computational error}} + \underbrace{\frac{1}{1-\kappa} \max \left\{ \frac{4\sqrt{5}}{C_3 w_{\min}^*} \eta\left(\frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s\right), \frac{\tilde{C}}{\sqrt{n}} \right\}}_{\text{statistical error}}, \end{aligned} \quad (9)$$

where the contraction coefficient  $\kappa = (\kappa_1 + \kappa_2)\kappa_3 \in (0, 1)$ , with  $\kappa_1 := 4\sqrt{5}w_{\max}^*C_3\epsilon/w_{\min}^*$ ,  $\kappa_2 := 2C_2\eta(\bar{\mathcal{E}}, s)/C_3^2$ , and  $\kappa_3 := 2/w_{\min}^* + 6\sqrt{2}$ . Here  $C_1, C_2, C_3$  are constants as defined in Assumption 1, and  $\tilde{C}$  is some positive constant.

Theorem 1 reveals the interplay between the computational error and the statistical error for our estimator, and has some interesting implications. First, it provides a theoretical guidance to terminate the iterations of the alternating updating algorithm. That is, when the computational error is dominated by the statistical error, in that

$$t \geq T := \log_{1/\kappa} \left( \frac{(1 - \kappa)\epsilon}{\max \left\{ \frac{4\sqrt{5}}{C_3 w_{\min}^*} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right), \frac{\tilde{C}}{\sqrt{n}} \right\}} \right),$$

the estimator from our algorithm achieves an error at the rate

$$O_p \left( \max \left\{ \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right), \frac{1}{\sqrt{n}} \right\} \right). \quad (10)$$

Second, Theorem 1 also provides a general theoretical guarantee for our estimator, for any distribution of the error tensor  $\mathcal{E}_i$ . The error rate depends on the distribution of the error tensor only through  $\eta(\frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s)$ . In the next two sections, we obtain the explicit forms of the error rates for some special error distributions.

We also remark that, Theorem 1 is obtained assuming the true rank  $K$  is known. This is a common assumption in theoretical analysis in the tensor literature (Zhou et al., 2013; Sun et al., 2016; Li and Zhang, 2016). To our knowledge, none of the exiting works has provided a provable estimation of the rank, due to that the exact tensor rank calculation is NP hard (Kolda and Bader, 2009). The theory of rank estimation will be our future research.

### 4.3 Gaussian tensor error

We next derive the explicit form of the statistical error in Theorem 1, when  $\mathcal{E}_i, i = 1, \dots, n$ , are i.i.d. Gaussian tensors; i.e., the entries of  $\mathcal{E}_i$  are i.i.d. standard Gaussian random variables.

**Corollary 1.** *Under the assumptions of Theorem 1, and assuming  $\mathcal{E}_i \in \mathbb{R}^{d_1 \times d_2 \times d_3}, i = 1, \dots, n$ , are i.i.d. Gaussian tensors, we have*

$$\eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) = O_p \left( \sqrt{\frac{s^3 \log(d_1 d_2 d_3)}{n}} \right).$$

Consequently the estimator at the  $T$ -th iteration of Algorithm 1 satisfies

$$\max \left\{ \max_k \|\widehat{w}_k^{(T)} - w_k^*\|_2, \max_{k,j} D(\widehat{\beta}_{k,j}^{(T)}, \beta_{k,j}^*) \right\} = O_p \left( \sqrt{\frac{s^3 \log(d_1 d_2 d_3)}{n}} \right).$$

To gain some insight about the statistical error  $O_p(\sqrt{s^3 \log(d_1 d_2 d_3)/n})$ , we note that,  $s$  is the maximal cardinality of vectors  $\widehat{\beta}_{k,j}$  for  $k \in [K]$  and  $j \in [m]$ . Since there are at most  $O(s^3)$  non-zero elements and  $O(d_1 d_2 d_3)$  free parameters in the true tensor coefficient  $\mathcal{B}^*$ , this rate is optimal. When the mode  $m = 1$ , it reduces to the statistical error  $O_p(\sqrt{s \log d/n})$  for the high-dimensional vector regression  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  with  $\beta \in \mathbb{R}^d$  and  $\|\beta\|_0 \leq s$ . This rate is known to be minimax optimal (Wang et al., 2014b; Yi and Caramanis, 2015).

#### 4.4 Symmetric matrix error

Next we consider the case when the response  $\mathcal{Y}_i$  is a symmetric matrix ( $m = 2$ ). Such a scenario is often encountered in functional neuroimaging analysis, where the target of interest is the brain connectivity pattern, encoded in the form of a symmetric correlation matrix. In this scenario, the symmetry of the coefficient  $\mathcal{B}^*$  requires that  $\beta_{k,1}^* = \beta_{k,2}^* = \beta_k^*$ . Henceforth,

$$\mathcal{B}^* = \sum_{k \in [K]} w_k^* \cdot \beta_k^* \circ \beta_k^* \circ \beta_{k,3}^*,$$

where  $w_k^* \in \mathbb{R}$ ,  $\|\beta_k^*\|_2 = 1$ ,  $\|\beta_k^*\|_0 \leq d_0$ , and  $\beta_{k,3}^* \in \mathbb{S}^p$ .

To facilitate the derivation of the explicit form of  $\eta(\frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s)$ , we assume that the error matrix  $\mathcal{E}_i$  satisfies

$$\mathcal{E}_i = (\widetilde{\mathcal{E}}_i + \widetilde{\mathcal{E}}_i^\top)/2, \tag{11}$$

where  $\widetilde{\mathcal{E}}_i \in \mathbb{R}^{d \times d}$  is a matrix whose entries are i.i.d. standard Gaussian. This assumption is mainly for technical reasons, and the theoretical analysis of a more general symmetric tensor response is left as future work.

**Corollary 2.** *Under the assumptions of Theorem 1, and assuming  $\mathcal{E}_i, i = 1, \dots, n$  are i.i.d. and of the form (11), we have*

$$\eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) = O_p \left( \sqrt{\frac{s^2 \log(d^2)}{n}} \right).$$

## 5 Simulations

In this section, we investigate the numerical performance of our method, and demonstrate its superior performance when compared to some alternative solutions.

To evaluate the estimation accuracy, we report the estimation error  $\|\widehat{\mathcal{B}} - \mathcal{B}\|_F$ . To evaluate the variable selection accuracy, we compute the true positive rate,  $\text{TPR} := m^{-1} \sum_{j=1}^m \text{TPR}_j$ , and the false positive rate,  $\text{FPR} := m^{-1} \sum_{j=1}^m \text{FPR}_j$ , where

$$\begin{aligned} \text{TPR}_j &:= \frac{1}{K} \sum_{k=1}^K \frac{\sum_i 1([\beta_{k,j}]_i \neq 0, [\widehat{\beta}_{k,j}]_i \neq 0)}{\sum_i 1([\beta_{k,j}]_i \neq 0)}, \\ \text{FPR}_j &:= \frac{1}{K} \sum_{k=1}^K \frac{\sum_i 1([\beta_{k,j}]_i = 0, [\widehat{\beta}_{k,j}]_i \neq 0)}{\sum_i 1([\beta_{k,j}]_i = 0)}. \end{aligned}$$

### 5.1 3D general tensor response example

We first consider an example of a third-order tensor response ( $m = 3$ ). The data is generated from model (1) with  $x_i$  a scalar taking values 0 or 1 with equal probability, and with the coefficient tensor  $\mathcal{B} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$  of the form,

$$\mathcal{B} = \sum_{k \in [K]} w_k \beta_{k,1} \circ \beta_{k,2} \circ \beta_{k,3} \circ \beta_{k,4}, \quad w_k \in \mathbb{R}, \quad \beta_{k,j} \in \mathbb{S}^{d_j}.$$

For each  $k \in [K]$ , we first generate i.i.d. standard Gaussian entries of the three vectors  $\beta_{k,1}, \beta_{k,2}, \beta_{k,3}$ , then truncate each vector with cardinality parameters  $s_{01}, s_{02}, s_{03}$  accordingly. Next we normalize each vector and aggregate the coefficients as  $w_k$ . We set  $\beta_{k,4}$  as one. In all simulations, we fix the sample size  $n = 20$ , and the tensor dimension  $d_1 = 100, d_2 = 50, d_3 = 20$ . We set the true cardinality  $s_{0j} = s * d_j$  ( $j = 1, 2, 3$ ) and vary  $s \in \{0.3, 0.5, 0.7\}$  to reflect different sparsity levels. We also vary the value of the rank  $K \in \{2, 5, 8\}$ . For parameter tuning, we set  $K \in \mathcal{K} = \{1, 2, \dots, 10\}$ , and  $s \in \mathcal{S} = \{0.1, 0.2, \dots, 0.9\}$ . We compare our method with two alternative solutions: the ordinary least squares (OLS) estimator that fits one response variable at a time and totally ignores the tensor structure, and the recent envelope-based tensor response regression (Li and Zhang, 2016).

Table 1 reports the average estimation error, and the true and false positive rates based on 20 replications. It is clearly seen that our method achieves a better performance than the



OLS and the envelope method, in terms of both estimation accuracy and variable selection accuracy. Such a pattern holds true for different sparsity levels (with  $s$  ranging from 0.3 to 0.7) and different rank values (with  $K$  ranging from 2 to 8). It is also noted that neither OLS nor the envelope method incorporates entry-wise sparsity, so TPR and FPR of both methods are always one.

## 5.2 2D symmetric matrix response example

Next we consider an example of a second-order symmetric tensor (matrix,  $m = 2$ ) example. Such an scenario is commonly encountered in functional neuroimaging and brain connectivity analysis. The data is generated from the following model

$$\mathcal{Y}_i = \mathcal{B} \times_{m+1} x_i + \mathcal{E}_i,$$

where  $\mathcal{Y}_i \in \mathbb{R}^{64 \times 64}$  is the symmetric matrix response,  $x_i$  is a scalar taking values 0 or 1, and the error term  $\mathcal{E}_i$  is a symmetric matrix whose upper triangle entries are generated randomly from a standard normal distribution. The sample size is  $n = 20$ . To mimic some common connectivity patterns, we consider three symmetric coefficient matrices  $\mathcal{B}$  that lead to three graph patterns: a randomly generated sparse and low-rank graph (random), a hub graph (hub), and a small world graph (small world). Note that  $\mathcal{B}$  is symmetric, though the resulting graph pattern is not necessarily so. The random and hub graphs are both low-rank, with the true rank  $K = 2$  and  $K = 10$ , respectively. The small world graph is not of an exact low-rank structure, and our method can be viewed as a low-rank *approximation*. It also allows us to investigate the performance of our method under potential model mis-specification.

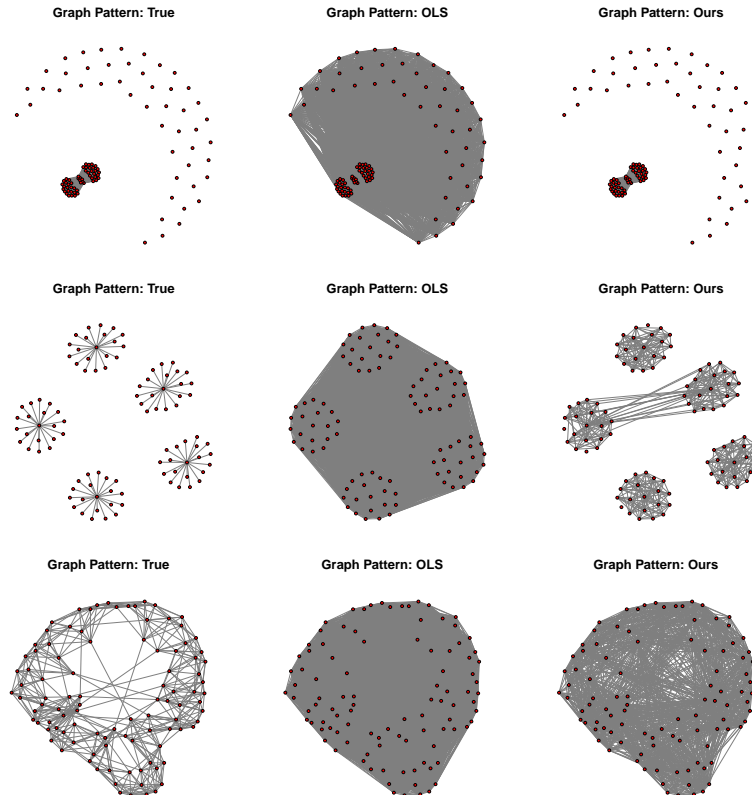
In Figure 1, the first column shows the three true graphs, the second column shows the estimates from the OLS for one data replication, and the third column shows those from our method for the same data. Note that the envelope method of [Li and Zhang \(2016\)](#) can not handle the symmetric response, and is thus excluded from this comparison. It is seen that, when the true graph is of a low-rank structure, such as the low-rank random graph and the hub graph, our method performs very well, and substantially outperforms the OLS solution, which generates a dense graph estimate in both cases. Even when the true graph is not an exactly low-rank, such as the small world graph, our method still manages to identify

Table 1: Third-order general tensor response example. Reported are the average estimation error, the true positive rate (TPR), and the false positive rate (FPR), with the standard error shown in the parenthesis. Three methods are compared: OLS, the envelope method (Env), and our approach (Ours), under different rank  $K$  and sparsity level  $s$ .

Rank	Sparsity	Method	Estimation Error	TPR	FPR
$K = 2$	$s = 0.3$	OLS	107.73 (2.70)	1 (0)	1 (0)
		Env	8.51 (0.07)	1 (0)	1 (0)
		Ours	3.37 (0.11)	1 (0)	0 (0)
	$s = 0.5$	OLS	107.73 (2.70)	1 (0)	1 (0)
		Env	8.55 (0.16)	1 (0)	1 (0)
		Ours	4.38 (0.13)	1 (0)	0 (0)
	$s = 0.7$	OLS	107.73 (2.70)	1 (0)	1 (0)
		Env	8.52 (0.12)	1 (0)	1 (0)
		Ours	5.24 (0.15)	1 (0)	0 (0)
$K = 5$	$s = 0.3$	OLS	103.05 (2.22)	1 (0)	1 (0)
		Env	13.68 (0.13)	1 (0)	1 (0)
		Ours	5.15 (0.12)	1 (0)	0 (0)
	$s = 0.5$	OLS	98.79 (2.25)	1 (0)	1 (0)
		Env	13.70 (0.13)	1 (0)	1 (0)
		Ours	6.38 (0.16)	1 (0)	0 (0)
	$s = 0.7$	OLS	101.49 (2.13)	1 (0)	1 (0)
		Env	13.87 (0.14)	1 (0)	1 (0)
		Ours	7.78 (0.16)	1 (0)	0 (0)
$K = 8$	$s = 0.3$	OLS	103.05 (2.22)	1 (0)	1 (0)
		Env	18.69 (0.17)	1 (0)	1 (0)
		Ours	6.43 (0.16)	1 (0)	0 (0)
	$s = 0.5$	OLS	99.93 (2.70)	1 (0)	1 (0)
		Env	18.87 (0.24)	1 (0)	1 (0)
		Ours	8.15 (0.21)	1 (0)	0 (0)
	$s = 0.7$	OLS	106.11 (2.63)	1 (0)	1 (0)
		Env	18.75 (0.28)	1 (0)	1 (0)
		Ours	10.48 (0.33)	1 (0)	0.04 (0.03)

a graph pattern that is closer to the truth than the OLS, which has a much more inflated false positive rate. Table 2 further reports the estimation error, the true and false positive rates over 20 data replications. The results reinforce the pattern we observe in Figure 1, and

Figure 1: Symmetric matrix response example. Shown are three graph patterns. Top row: the random graph, middle row: the hub graph, bottom row: the small world graph. Left column: the true graph pattern, middle column: the estimated graph pattern from OLS, right column: the estimate from our method.



shows the superior performance of our method.

## 6 Real data analysis

In this section, we analyze a real neuroimaging data to illustrate our proposed method. The data is from the Autism Brain Imaging Data Exchange (ABIDE), a study for autism spectrum disorder (ASD) (Di Martino et al., 2014). ASD is an increasingly prevalent neurodevelopmental disorder. It is characterized by symptoms such as social difficulties, communication deficits, stereotyped behaviors and cognitive delays (Rudie et al., 2013). After removing those with missing values and poor image quality, the ABIDE study consists of the resting-state functional magnetic resonance imaging (fMRI) of 795 subjects, of which

Table 2: Symmetric matrix response example. Reported are the average estimation error, the true positive rate (TPR), and the false positive rate (FPR), with the standard error shown in the parenthesis. Two methods are compared: OLS and our approach (Ours), for three different graph patterns.

Graph Pattern	Method	Estimation Error	TPR	FPR
Random	OLS	22.17 (0.56)	1 (0)	1 (0)
	Ours	2.50 (0.08)	1 (0)	0 (0)
Hub	OLS	22.17 (0.56)	1 (0)	1 (0)
	Ours	7.20 (0.44)	0.83 (0.04)	0.08 (0.00)
Small World	OLS	22.17 (0.56)	1 (0)	1 (0)
	Ours	17.26 (0.12)	0.91 (0.03)	0.21 (0.02)

362 have ASD, and 433 are normal controls. It is of scientific interest to understand how brain functional architecture and brain connectivity pattern differ between subjects with ASD and normal controls. For each individual subject, there are two forms of data summary of the fMRI scan. The first is fractional amplitude of low-frequency fluctuations (fALFF). It is a metric reflecting the percentage of power spectrum within low-frequency domain (0.01 – 0.1 Hz). It characterizes the intensity of spontaneous brain activities, and thus provides a measure of functional architecture of the brain (Shi and Kang, 2015). The second is the partial correlation between brain regions (Peng et al., 2009; Wang et al., 2016). It reveals the synchronization of brain systems and offers an alternative measure of the intrinsic brain functional architecture. It is noted that the fALFF and the partial correlation are measured at different scales. The fALFF is calculated at each individual image voxel, and forms a third-order tensor of dimension  $91 \times 109 \times 91$ . The partial correlation is computed between pairs of pre-specified brain regions, each of which is a collection of brain voxels. It forms a symmetric matrix of dimension  $116 \times 116$ , corresponding to 116 regions-of-interest under the commonly used Anatomical Automatic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). We fit two tensor response regression models. One takes the fALFF tensor as a response, and the other takes the symmetric partial correlation matrix as a response. The covariate in our analysis is the ASD status indicator (1 = ASD and 0 = normal control).

Figure 2 reports the results of the regression with the third-order fALFF tensor response.

Table 3: Analysis of the ABIDE data. The response is the third-order fALFF tensor. Reported are the identified brain regions by our method that exhibit clear difference in fALFF between the ASD and normal control.

Important Regions	$\widehat{\mathcal{B}}_{i,j,k}$
Postcentral_L	0.0033
Cerebellum_9_L	0.0028
Precuneus_L	0.0025
Occipital_Sup_R	0.0025
Parietal_Inf_L	0.0025
Cerebellum_8_L	0.0020
Cerebellum_8_R	0.0020
Precuneus_R	0.0017
Parietal_Sup_R	0.0017
Cerebellum_9_R	0.0016
Parietal_Sup_L	0.0016

Shown in the plot are the estimated coefficient tensors from OLS and our method, overlaid on a brain image of a randomly selected subject. The OLS essentially concludes the entire brain regions are different between the ASD group and the control. By contrast, our method identifies a small number of brain regions that exhibit clear difference between the two groups. After we obtain the estimated coefficient tensor  $\widehat{\mathcal{B}}$ , we map those nonzero entries to the AAL atlas. Table 3 reports the names of those identified regions and the corresponding coefficient entries in each region. Our results in general agree with the autism literature. For instance, we have found that multiple cerebellum (**Cerebellum**) regions show distinctive patterns between the ASD and control groups. The cerebellum has been long known for its importance in motor learning, coordination, and more recently, cognitive functions and affective regulation. It has emerged as one of the key brain regions affected in autism ([Becker and Stoodley, 2013](#)). Furthermore, we identified the superior parietal lobule (**Parietal\_Sup**) and precuneus (**Precuneus**), which agrees with [Travers et al. \(2015\)](#), in that they found individuals with ASD showed decreased activation in the superior parietal lobule and precuneus relative to individuals with typical development, suggesting that the superior parietal lobule may play an important role in motor learning and repetitive behavior in individuals with ASD.

Figure 2: Analysis of the ABIDE data. The response is a third-order brain image tensor. Shown are the estimated coefficient tensor overlaid on a randomly selected brain image. Top row: OLS, bottom row: our method. Left column: front view, middle column: side view, right column: top view.

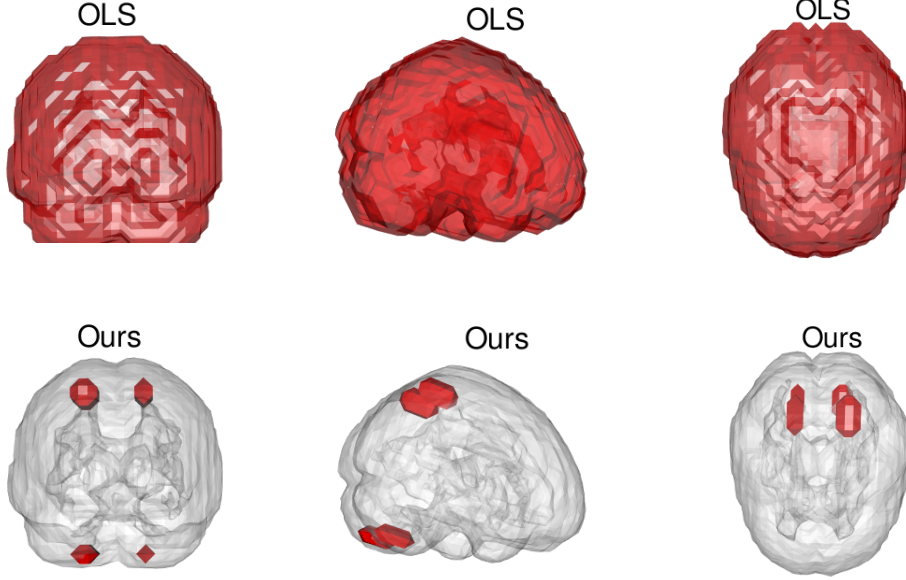
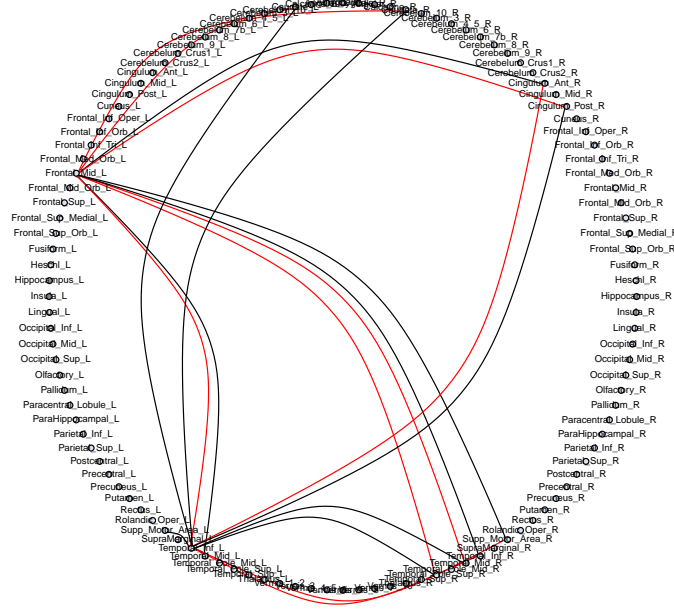


Figure 3 reports the top 20 identified links from our method using the second-order symmetric partial correlation matrix as a response. The red links correspond to those that are more likely to be absent in the ASD group, whereas the black ones are those more likely to be present in ASD. Table 4 reports the names of those links and their relative connection strengthens in the network, as reflected by the corresponding entries in the matrix coefficient  $\hat{\mathcal{B}}$ . Many different connectivity patterns concentrate on the left middle frontal gyrus (`Frontal_Mid_L`) and the temporal lobe (`Temporal`), and such findings again generally agrees with the autism literature ([Kana et al., 2006](#); [Di Martino et al., 2014](#); [Ha et al., 2015](#)).

## Appendix

In this appendix, we first introduce a number of auxiliary lemmas and then provide detailed proofs of the main theorem. The technical proofs of all lemmas and corollaries are provided in the online supplement.

Figure 3: Analysis of the ABIDE data. The response is a symmetric partial correlation matrix. Shown are the top 20 links found by our method. Red links are more likely to be absent in the ASD group and black link are more likely to be present in the ASD group.



## A1 Auxiliary lemmas

**Lemma 3.** For any matrix  $\mathbf{Y}_i \in \mathbb{R}^{d_1 \times d_2}$ , with  $i = 1, \dots, n$ , and vectors  $\gamma = (\gamma_1, \dots, \gamma_n)^\top \in \mathbb{R}^n$ ,  $\alpha \in \mathbb{R}^{d_1}$  and  $\beta \in \mathbb{R}^{d_2}$ , we have

$$\arg \min_{\beta, \|\beta\|_2=1} \frac{1}{n} \sum_{i=1}^n \gamma_i^2 \left\| \mathbf{Y}_i - \alpha \beta^\top \right\|_F^2 = \arg \min_{\beta, \|\beta\|_2=1} \left\| \frac{1}{n} \sum_{i=1}^n \gamma_i^2 \mathbf{Y}_i - \alpha \beta^\top \right\|_F^2.$$

The proof of Lemma 3 is provided in Section S.3 in the online supplement.

We next introduce the Slepian's lemma (Slepian, 1962), which provides a comparison between the supremums of two Gaussian processes.

**Lemma 4.** (Slepian, 1962, Slepian's Lemma) Denote two centered Gaussian processes  $\{G_s, s \in \mathcal{S}\}$  and  $\{H_s, s \in \mathcal{S}\}$ . Assume that both processes are almost surely bounded and for each  $s, t \in \mathcal{S}$ ,  $\mathbb{E}(G_s - G_t)^2 \leq \mathbb{E}(H_s - H_t)^2$ , then we have

$$\mathbb{E} \left[ \sup_{s \in \mathcal{S}} G_s \right] \leq \mathbb{E} \left[ \sup_{s \in \mathcal{S}} H_s \right].$$

Table 4: Analysis of the ABIDE data. The response is a symmetric partial correlation matrix. Reported are the identified top 20 links by our method.

Connected Regions	$\widehat{\mathcal{B}}_{i,j}$
Frontal_Mid_L $\longleftrightarrow$ Temporal_Inf_L	-0.0034
Cingulum_Ant_R $\longleftrightarrow$ Temporal_Inf_L	-0.0027
Temporal_Inf_L $\longleftrightarrow$ Temporal_Inf_R	-0.0024
Supp_Motor_Area_R $\longleftrightarrow$ Temporal_Inf_L	-0.0023
Temporal_Mid_L $\longleftrightarrow$ Temporal_Inf_L	-0.0023
Frontal_Mid_L $\longleftrightarrow$ Temporal_Mid_R	-0.0020
Frontal_Mid_L $\longleftrightarrow$ Caudate_R	-0.0019
Frontal_Mid_L $\longleftrightarrow$ Temporal_Pole_Sup_R	-0.0018
Frontal_Mid_L $\longleftrightarrow$ Cingulum_Post_R	-0.0017
Frontal_Mid_L $\longleftrightarrow$ Caudate_L	-0.0016
Frontal_Mid_L $\longleftrightarrow$ Supp_Motor_Area_R	0.0017
Frontal_Mid_L $\longleftrightarrow$ Temporal_Mid_L	0.0017
Frontal_Mid_L $\longleftrightarrow$ Temporal_Inf_R	0.0019
Frontal_Mid_L $\longleftrightarrow$ Cingulum_Ant_R	0.0021
Supp_Motor_Area_L $\longleftrightarrow$ Temporal_Inf_L	0.0021
Caudate_L $\longleftrightarrow$ Temporal_Inf_L	0.0021
Cingulum_Post_R $\longleftrightarrow$ Temporal_Inf_L	0.0023
Temporal_Pole_Sup_R $\longleftrightarrow$ Temporal_Inf_L	0.0024
Caudate_R $\longleftrightarrow$ Temporal_Inf_L	0.0026
Temporal_Mid_R $\longleftrightarrow$ Temporal_Inf_L	0.0026

Moreover, if  $\mathbb{E}(G_s^2) = E(H_s^2)$  for all  $s \in \mathcal{S}$ , then we have, for each  $x > 0$ ,

$$\mathbb{P} \left[ \sup_{s \in \mathcal{S}} G_s > x \right] \leq \mathbb{P} \left[ \sup_{s \in \mathcal{S}} H_s > x \right].$$

The next result provides a concentration of Lipschitz functions of Gaussian random variables (Massart, 2003).

**Lemma 5.** (Massart, 2003, Theorem 3.4) Let  $\mathbf{v} \in \mathbb{R}^d$  be a Gaussian random variable such that  $\mathbf{v} \sim N(0, \mathbf{I}_d)$ . Assuming  $g(\mathbf{v}) \in \mathbb{R}$  to be a Lipschitz function such that  $|g(\mathbf{v}_1) - g(\mathbf{v}_2)| \leq L \|\mathbf{v}_1 - \mathbf{v}_2\|_2$  for any  $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$ , then we have, for each  $t > 0$ ,

$$\mathbb{P} [|g(\mathbf{v}) - \mathbb{E}[g(\mathbf{v})]| \geq t] \leq 2 \exp \left( -\frac{t^2}{2L^2} \right).$$



The next lemma provides an upper bound of the Gaussian width of the unit ball for the sparsity regularizer (Raskutti and Yuan, 2016).

**Lemma 6.** (Raskutti and Yuan, 2016, Lemma 2) For a tensor  $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ , denote its regularizer  $R(\mathcal{T}) = \sum_{j_1} \sum_{j_2} \sum_{j_3} |\mathcal{T}_{j_1, j_2, j_3}|$ . Define the unit ball of this regularizer as  $B_R(1) := \{\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3} | R(\mathcal{T}) \leq 1\}$ . For a Gaussian tensor  $\mathcal{G} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$  whose entries are independent standard normal random variables, we have

$$\mathbb{E} \left[ \sup_{\mathcal{T} \in B_R(1)} \langle \mathcal{T}, \mathcal{G} \rangle \right] \leq c \sqrt{\log(d_1 d_2 d_3)},$$

for some bounded constant  $c > 0$ .

The next lemma links the hard thresholding sparsity and the  $L_1$ -penalized sparsity.

**Lemma 7.** For any vectors  $\mathbf{u} \in \mathbb{R}^{d_1}, \mathbf{v} \in \mathbb{R}^{d_2}, \mathbf{w} \in \mathbb{R}^{d_3}$  satisfying  $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1, \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s$ , and  $\|\mathbf{w}\|_0 \leq s$ , denoting  $\mathcal{A} := \mathbf{u} \circ \mathbf{v} \circ \mathbf{w}$ , we have the bound of the  $L_1$ -norm regularizer

$$\|\mathcal{A}\|_1 := \sum_{j_1} \sum_{j_2} \sum_{j_3} |\mathcal{A}_{j_1 j_2 j_3}| \leq s^{3/2}.$$

*Proof:* According to the Cauchy-Schwarz inequality, we have  $\|\mathbf{u}\|_1 \leq \sqrt{s} \|\mathbf{u}\|_2 = \sqrt{s}$ , and  $\|\mathbf{v}\|_1 \leq \sqrt{s}, \|\mathbf{w}\|_1 \leq \sqrt{s}$ . Therefore,  $\|\mathcal{A}\|_1 = \|\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}\|_1 \leq \|\mathbf{u}\|_1 \cdot \|\mathbf{v}\|_1 \cdot \|\mathbf{w}\|_1 \leq s^{3/2}$ .  $\square$

## A2 Proof of Theorem 1

We divide the proof of Theorem 1 into two major steps: characterization of the estimation error in Step 1 in Algorithm 1, then the estimation error in Step 2. Each leads to a new theorem. Then we complete the proof of Theorem 1 by iteratively applying those results.

### A2.1 Estimation error in Step 1 in Algorithm 1

We first derive the estimation error in Step 1 of our Algorithm 1. The key idea is to transform the problem in Step 1 into a standard sparse tensor decomposition problem, then incorporate the existing contracting results obtained in Sun et al. (2016) to derive the final error bound of the estimator in Step 1.

In the following derivation, for simplicity, we assume  $K = 1$ . This does not lose generality, since  $K$  is assumed to be a constant, and it does not affect the final error rate. In this case, the true model reduces to  $\mathcal{Y}_i = w^*(\beta_{k,m+1}^{*\top} \mathbf{x}_i) \beta_{k,1}^* \circ \cdots \circ \beta_{k,m}^* + \mathcal{E}_i$ , for each  $i = 1, \dots, n$ . Based on the Step 1 of our algorithm, if the true parameter  $\beta_{k,m+1}^*$  is available, we are solving the following sparse tensor decomposition problem,

$$\bar{\mathcal{R}}_k = \mathcal{T} + \bar{\mathcal{E}}, \quad (12)$$

where the true tensor  $\mathcal{T} = w_k^* \beta_{k,1}^* \circ \cdots \circ \beta_{k,m}^*$ , the oracle response and the oracle error are, respectively,

$$\bar{\mathcal{R}}_k := \frac{1}{n} \sum_{i=1}^n \mathcal{R}_{ik} = \frac{1}{n} \sum_{i=1}^n \frac{\mathcal{Y}_i}{\beta_{k,m+1}^{*\top} \mathbf{x}_i} \text{ and } \bar{\mathcal{E}} = \frac{1}{n} \sum_{i=1}^n \frac{\mathcal{E}_i}{\beta_{k,m+1}^{*\top} \mathbf{x}_i}.$$

In practice, however, we only have an estimator  $\hat{\beta}_{k,m+1}$ . Hence the sparse tensor decomposition method in Step 1 of our algorithm is actually applied to

$$\hat{\mathcal{R}}_k = \mathcal{T} + \hat{\mathcal{E}} \quad (13)$$

with the response tensor

$$\hat{\mathcal{R}}_k := \frac{1}{n} \sum_{i=1}^n \frac{\mathcal{Y}_i}{\hat{\beta}_{k,m+1}^\top \mathbf{x}_i}$$

Therefore, according to (12) and (13), we have the explicit form of  $\hat{\mathcal{E}}$ ,

$$\begin{aligned} \hat{\mathcal{E}} &= \hat{\mathcal{R}}_k - \bar{\mathcal{R}}_k + \bar{\mathcal{E}} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{(\beta_{k,m+1}^* - \hat{\beta}_{k,m+1})^\top \mathbf{x}_i}{\hat{\beta}_{k,m+1}^\top \mathbf{x}_i \beta_{k,m+1}^{*\top} \mathbf{x}_i} \mathcal{Y}_i}_{I_1} + \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{\mathcal{E}_i}{\beta_{k,m+1}^{*\top} \mathbf{x}_i}}_{I_2} \end{aligned} \quad (14)$$

Before we derive the estimation error of the estimator based on (13), we introduce a lemma (Sun et al., 2016) for deriving the error bound of a general sparse tensor decomposition.

**Assumption 4.** *The decomposition components are incoherent such that*

$$\zeta := \max_{i \neq j} \{|\langle \beta_{i,1}^*, \beta_{j,1}^* \rangle|, |\langle \beta_{i,2}^*, \beta_{j,2}^* \rangle|, |\langle \beta_{i,3}^*, \beta_{j,3}^* \rangle|\} \leq \frac{C_0}{\sqrt{d_0}}, \quad (15)$$

with  $d_0 = \max\{d_{01}, d_{02}, d_{03}\}$ , and for any  $j$ ,  $\|\sum_{i \neq j} w_i \langle \beta_{i,1}^*, \beta_{j,1}^* \rangle \langle \beta_{i,2}^*, \beta_{j,2}^* \rangle \beta_{i,3}^*\| \leq C_1 w_{\max}^* \sqrt{K} \zeta$ . Moreover, the matrices  $\mathbf{A} := [\beta_{i,1}^*, \dots, \beta_{K,1}^*]$ ,  $\mathbf{B} := [\beta_{i,2}^*, \dots, \beta_{K,2}^*]$ , and  $\mathbf{C} := [\beta_{i,3}^*, \dots, \beta_{K,3}^*]$  satisfy that  $\max\{\|\mathbf{A}\|, \|\mathbf{B}\|, \|\mathbf{C}\|\} \leq 1 + C_2 \sqrt{K/d_0}$  for some positive constants  $C_0, C_1, C_2$ .

Define a function  $f(\epsilon; K, d_0)$  as

$$f(\epsilon; K, d_0) := \frac{2C_0}{\sqrt{d_0}} \left(1 + C_2 \sqrt{\frac{K}{d_0}}\right)^2 \epsilon + C_1 \frac{\sqrt{K}}{d_0} + C_3 \epsilon^2, \quad (16)$$

for some constants  $C_0, C_1, C_2, C_3 > 0$ . When  $K = o(d_0^{3/2})$ , the first two terms of  $f(\epsilon; K, d_0)$  converge to 0 and the last term is the contracting term.

**Lemma 8.** (*Sun et al., 2016, Lemma S.4.1*) Consider the model  $\widehat{\mathcal{T}} = \mathcal{T} + \mathcal{E}$  where the low-rank and sparse components of  $\mathcal{T}$  satisfy Assumption 4, and assume  $\|\mathcal{T}\| \leq C_3 w_{\max}^*$  and  $K = o(d_0^{3/2})$ . In addition, assume the estimators  $\widehat{\beta}_{j,1}$  and  $\widehat{\beta}_{j,2}$  satisfy  $D(\widehat{\beta}_{j,1}, \beta_{j,1}^*) \leq \epsilon$  and  $D(\widehat{\beta}_{j,2}, \beta_{j,2}^*) \leq \epsilon$  for some  $j \in [K]$ . If the perturbation error  $\eta(\mathcal{E}, d_0 + s)$ , with  $s \geq d_0$ , is small enough such that  $\eta(\mathcal{E}, d_0 + s) < w_j(1 - \epsilon^2) - w_{\max}^* f(\epsilon; K, d_0)$ , then the update  $\widehat{\beta}_{j,3}$  satisfies, with high probability,

$$D(\widehat{\beta}_{j,3}, \beta_{j,3}^*) \leq \frac{\sqrt{5} w_{\max}^* f(\epsilon; K, d_0) + \sqrt{5} \eta(\mathcal{E}, d_0 + s)}{w_j(1 - \epsilon^2) - w_{\max}^* f(\epsilon; K, d_0) - \eta(\mathcal{E}, d_0 + s)}. \quad (17)$$

If we further assume  $D(\widehat{\beta}_{j,3}, \beta_{j,3}^*) \leq \epsilon$ , then the update  $\widehat{w} = \widehat{\mathcal{T}} \times_1 \widehat{\beta}_{j,1} \times_2 \widehat{\beta}_{j,2} \times_3 \widehat{\beta}_{j,3}$  satisfies, with high probability,  $|\widehat{w} - w_j| \leq 2w_j \epsilon^2 + w_{\max}^* f(\epsilon; K, d_0) + \eta(\mathcal{E}, d_0 + s)$ .

By verifying the conditions in Lemma 8, we are able to compute the estimation error in Step 1 of our algorithm. For simplicity, we consider the case when  $m = 3$ , while the extension to a more general  $m$  is straightforward.

**Lemma 9.** Assume  $\|\mathcal{T}\| \leq C_1 w_{\max}^*$ ,  $D(\widehat{\beta}_{k,1}, \beta_{k,1}^*) \leq \epsilon$ ,  $D(\widehat{\beta}_{k,2}, \beta_{k,2}^*) \leq \epsilon$ , and

$$\epsilon < \min \left\{ \sqrt{\frac{w_{\min}^*}{2(w_{\min}^* + w_{\max}^* C_1)}}, \frac{w_{\min}^*}{4\sqrt{5} w_{\max}^* C_1} \right\}.$$

Assume the error tensor satisfies  $\eta(\widehat{\mathcal{E}}, d_0 + s) \leq w_{\min}^*/4$  with  $s \geq d_0$ , where  $\widehat{\mathcal{E}}$  is as defined in (14). Then we have

$$D(\widehat{\beta}_{k,3}, \beta_{k,3}^*) \leq \kappa_1 \epsilon + \frac{4\sqrt{5}}{w_{\min}^*} \eta(\widehat{\mathcal{E}}, d_0 + s), \quad (18)$$

where the contraction coefficient

$$\kappa_1 := \frac{4\sqrt{5} w_{\max}^* C_1}{w_{\min}^*} \epsilon < \frac{4\sqrt{5} w_{\max}^* C_1}{w_{\min}^*} \min \left\{ \sqrt{\frac{w_{\min}^*}{2(w_{\min}^* + w_{\max}^* C_1)}}, \frac{w_{\min}^*}{4\sqrt{5} w_{\max}^* C_1} \right\} \in (0, 1).$$

The proof of Lemma 9 is provided in Section S.4 in the online supplement. Based on Lemma 9, to compute the closed-form error rate in Step 1 of our Algorithm, the remaining step is to compute  $\eta(\widehat{\mathcal{E}}, s)$ , since  $\eta(\widehat{\mathcal{E}}, d_0 + s) \leq 2\eta(\widehat{\mathcal{E}}, s)$  by noting that  $s \geq d_0$ . Here the explicit form of  $\widehat{\mathcal{E}}$  is defined in (14). Again we only consider  $m = 3$  for simplicity, and the proof for a general  $m$  follows straightforwardly.

**Lemma 10.** *Assume the conditions in Lemma 9 hold. Assume that  $\|w_k^* \beta_{k,1}^* \circ \dots \circ \beta_{k,m}^*\| \leq C_1$ ,  $\|\mathbf{x}_i\| \leq C_2$ , and  $|\beta_{k,m+1}^{*\top} \mathbf{x}_i| \geq C_3$  for each  $i = 1, \dots, n$ , for some positive constants  $C_1, C_2, C_3$ . If  $\epsilon \leq C_3/(2C_2)$ , then we have*

$$\eta(\widehat{\mathcal{E}}, s) \leq \underbrace{\frac{2C_2\eta(\bar{\mathcal{E}}, s)}{C_3^2}}_{\kappa_2} \epsilon + \frac{1}{C_3} \eta(\bar{\mathcal{E}}, s).$$

where  $\bar{\mathcal{E}} := \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i$ .

The proof of Lemma 10 is provided in Section S.5 in the online supplement. Combining Lemma 9 and Lemma 10, we obtain the final contraction result of Step 1 in Algorithm 1.

**Theorem 2.** *(Contraction result in Step 1 in Algorithm 1) Assume  $D(\widehat{\beta}_{k,1}, \beta_{k,1}^*) \leq \epsilon$ ,  $D(\widehat{\beta}_{k,2}, \beta_{k,2}^*) \leq \epsilon$ , and  $D(\widehat{\beta}_{k,4}, \beta_{k,4}^*) \leq \epsilon$ , with*

$$\epsilon < \min \left\{ \sqrt{\frac{w_{\min}^*}{2(w_{\min}^* + w_{\max}^* C_1)}}, \frac{w_{\min}^*}{4\sqrt{5}w_{\max}^* C_1}, \frac{C_3}{2C_2} \right\}.$$

Assume the error tensor satisfies  $\eta(n^{-1} \sum_{i=1}^n \mathcal{E}_i, d_0 + s) \leq w_{\min}^*/4$  with  $s \geq d_0$ . Then we have

$$D(\widehat{\beta}_{k,3}, \beta_{k,3}^*) \leq (\kappa_1 + \kappa_2)\epsilon + \frac{4\sqrt{5}}{C_3 w_{\min}^*} \eta\left(\frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s\right). \quad (19)$$

## A2.2 Estimation error in Step 2 in Algorithm 1

Next we derive the estimation error in Step 2 of our algorithm. That is, we aim to bound  $D(\widehat{\beta}_{k,m+1}, \beta_{k,m+1}^*)$  given the estimators  $\widehat{w}_k, \widehat{\beta}_{k,1}, \dots, \widehat{\beta}_{k,m}$ .

Denote  $\widehat{\mathcal{A}}_k = \widehat{w}_k \widehat{\beta}_{k,1} \circ \dots \circ \widehat{\beta}_{k,m}$ , and  $\widehat{\mathcal{T}}_i = \mathcal{Y}_i - \sum_{k' \neq k, k' \in [K]} \widehat{w}_{k'} (\widehat{\beta}_{k',m+1}^\top \mathbf{x}_i) \widehat{\beta}_{k',1} \circ \dots \circ \widehat{\beta}_{k',m}$ , and the closed-form estimator in Step 2 of our algorithm is

$$\widehat{\beta}_{k,m+1} = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \frac{n^{-1} \sum_{i=1}^n \langle \widehat{\mathcal{T}}_i, \widehat{\mathcal{A}}_k \rangle \mathbf{x}_i}{\|\widehat{\mathcal{A}}_k\|_F^2}.$$

**Theorem 3.** (Contraction result in Step 2 in Algorithm 1) Under Assumption 1, and the assumption that the initialization error satisfies  $\epsilon \leq w_{\min}^*/2$ , if  $|\hat{w}_k - w_k^*| \leq \epsilon$ ,  $D(\hat{\beta}_{k,1}, \beta_{k,1}^*) \leq \epsilon, \dots, D(\hat{\beta}_{k,m}, \beta_{k,m}^*) \leq \epsilon$ , then we have

$$D(\hat{\beta}_{k,m+1}, \beta_{k,m+1}^*) \leq \kappa_3 \epsilon + \frac{\tilde{C}}{\sqrt{n}}, \quad (20)$$

where  $\kappa_3 := 2/w_{\min}^* + 6\sqrt{2}$  and  $\tilde{C}$  is a positive constant as defined in (24).

*Proof:* For simplicity, we only prove for  $K = 1$  and  $m = 3$ . The derivation for a general  $K$  and  $m$  follows similarly.

Denote  $\mathcal{A}_k^* := w_k^* \beta_{k,1}^* \circ \beta_{k,2}^* \circ \beta_{k,3}^*$ . The true model reduces to  $\mathcal{Y}_i = (\beta_{k,4}^{*\top} \mathbf{x}_i) \mathcal{A}_k^* + \mathcal{E}_i$ , for each  $i = 1, \dots, n$ . Denote  $\Omega := (n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top)^{-1}$ . Given  $\hat{\mathcal{A}}_k = \hat{w}_k \hat{\beta}_{k,1} \circ \dots \circ \hat{\beta}_{k,3}$ , and  $\hat{\mathcal{T}}_i = \mathcal{Y}_i$  when  $K = 1$ , we have the following simplification of  $\hat{\beta}_{k,4}$ ,

$$\begin{aligned} \hat{\beta}_{k,4} &= \Omega \frac{\sum_{i=1}^n \langle \hat{\mathcal{T}}_i, \hat{\mathcal{A}}_k \rangle \mathbf{x}_i}{n \|\hat{\mathcal{A}}_k\|_F^2}, \\ &= \Omega \frac{\sum_{i=1}^n \langle (\beta_{k,4}^{*\top} \mathbf{x}_i) \mathcal{A}_k^* + \mathcal{E}_i, \hat{\mathcal{A}}_k \rangle \mathbf{x}_i}{n \|\hat{\mathcal{A}}_k\|_F^2}, \\ &= \frac{\langle \mathcal{A}_k^*, \hat{\mathcal{A}}_k \rangle}{\|\hat{\mathcal{A}}_k\|_F^2} \beta_{k,4}^* + \frac{\Omega \sum_{i=1}^n \langle \mathcal{E}_i, \hat{\mathcal{A}}_k \rangle \mathbf{x}_i}{n \|\hat{\mathcal{A}}_k\|_F^2}, \end{aligned} \quad (21)$$

where the first part in (21) is due to the fact that  $n^{-1} \sum_{i=1}^n (\beta_{k,4}^{*\top} \mathbf{x}_i) \Omega \mathbf{x}_i = n^{-1} \sum_{i=1}^n \Omega \mathbf{x}_i \mathbf{x}_i^\top \beta_{k,4}^* = \beta_{k,4}^*$ . Therefore, the error bound of  $\hat{\beta}_{k,4}$  can be simplified as

$$\begin{aligned} \left\| \hat{\beta}_{k,4} - \beta_{k,4}^* \right\|_2 &= \left\| \frac{\langle \mathcal{A}_k^*, \hat{\mathcal{A}}_k \rangle}{\|\hat{\mathcal{A}}_k\|_F^2} \beta_{k,4}^* - \beta_{k,4}^* + \frac{\Omega \sum_{i=1}^n \langle \mathcal{E}_i, \hat{\mathcal{A}}_k \rangle \mathbf{x}_i}{n \|\hat{\mathcal{A}}_k\|_F^2} \right\|_2, \\ &\leq \underbrace{\left| \frac{\langle \mathcal{A}_k^*, \hat{\mathcal{A}}_k \rangle}{\|\hat{\mathcal{A}}_k\|_F^2} - 1 \right|}_{(I)} \|\beta_{k,4}^*\|_2 + \underbrace{\left\| \frac{\Omega \sum_{i=1}^n \langle \mathcal{E}_i, \hat{\mathcal{A}}_k \rangle \mathbf{x}_i}{n \|\hat{\mathcal{A}}_k\|_F^2} \right\|_2}_{(II)}. \end{aligned} \quad (22)$$

In the following lemma, we bound two terms in (22) to obtain the final error bound.

**Lemma 11.** Under the Conditions in Theorem 3, we have

$$(I) \leq \left[ \frac{2}{w_{\min}^*} + 6\sqrt{2} \right] \epsilon, \quad (23)$$

$$(II) \leq \underbrace{\frac{\max_i \|\Omega \mathbf{x}_i\|_2 \cdot \mathbb{E}[\|\mathcal{G}\|_F]}{\|\hat{\mathcal{A}}_k\|_F}}_{\tilde{C}} \cdot \frac{1}{\sqrt{n}}. \quad (24)$$

Finally, combining the results in (23) and (24) leads to the final bound of  $\|\widehat{\beta}_{k,4} - \beta_{k,4}^*\|_2$ , and hence the bound of  $D(\widehat{\beta}_{k,4}, \beta_{k,4}^*)$ .  $\square$

### A2.3 Proof of Theorem 1

Now we complete the proof of Theorem 1. It follows by iteratively applying the contraction results in Theorem 2 for Step 1 of Algorithm 1, and Theorem 3 for Step 2 of Algorithm 1.

In iteration  $t = 1$ , given the initializations  $\widehat{\beta}_{k,j}^{(0)}$  and  $\widehat{w}^{(0)}$  with initialization error  $\epsilon$ , Theorem 2 implies that

$$D(\widehat{\beta}_{k,3}^{(1)}, \beta_{k,3}^*) \leq (\kappa_1 + \kappa_2)\epsilon + \frac{4\sqrt{5}}{C_3 w_{\min}^*} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right),$$

where  $\kappa_1 + \kappa_2 < 1$  according to Assumptions 2 and 3. The second term converges to zero as sample size increases. Therefore, for a sufficiently large sample size, the above error bound is smaller than  $\epsilon$ . By a similar derivation, the same error bound holds for  $D(\widehat{\beta}_{k,1}^{(1)}, \beta_{k,1}^*)$ ,  $D(\widehat{\beta}_{k,2}^{(1)}, \beta_{k,2}^*)$ , and  $|\widehat{w}_k^{(1)} - w_k^*|$ . In Step 2 of the algorithm, applying Theorem 3 based on the above estimators in iteration  $t = 1$ , we obtain that

$$D(\widehat{\beta}_{k,4}^{(1)}, \beta_{k,4}^*) \leq \kappa_3(\kappa_1 + \kappa_2)\epsilon + \kappa_3 \frac{4\sqrt{5}}{C_3 w_{\min}^*} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) + \frac{\widetilde{C}}{\sqrt{n}}.$$

Again, the contraction coefficient  $\kappa_3(\kappa_1 + \kappa_2) < 1$  according to Assumptions 2 and 3, and the remaining term converges to zero as sample size increases. For  $\kappa = (\kappa_1 + \kappa_2)\kappa_3 \in (0, 1)$ , by repeatedly applying these derivations, in the  $t$  iteration, we obtain that

$$\begin{aligned} & \max \left\{ \max_k \|\widehat{w}_k^{(t)} - w_k^*\|_2, \max_{k,j} \{\|\widehat{\beta}_{k,j}^{(t)} - \beta_{k,j}^*\|_2\} \right\} \\ & \leq \kappa^t \epsilon + \frac{1 - \kappa^t}{1 - \kappa} \frac{4\sqrt{5}}{C_3 w_{\min}^*} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) + \frac{1 - \kappa^{t-1}}{1 - \kappa} \frac{\widetilde{C}}{\sqrt{n}} \\ & \leq \kappa^t \epsilon + \frac{1}{1 - \kappa} \max \left\{ \frac{4\sqrt{5}}{C_3 w_{\min}^*} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right), \frac{\widetilde{C}}{\sqrt{n}} \right\}. \end{aligned}$$

This completes the proof of Theorem 1.  $\square$

## References

- ANANDKUMAR, A., GE, R. and JANZAMIN, M. (2014). Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180* .
- BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2016). Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics* To appear.
- BECKER, E. B. and STOODLEY, C. J. (2013). Chapter one - autism spectrum disorder and the cerebellum. In *Neurobiology of Autism* (G. Konopka, ed.), vol. 113 of *International Review of Neurobiology*. Academic Press, 1 – 34.
- CHI, E. C. and KOLDA, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications* **33** 1272–1299.
- DI MARTINO, A., YAN, C.-G., LI, Q., DENIO, E., CASTELLANOS, F. X., ALAERTS, K., ANDERSON, J. S., ASSAF, M., BOOKHEIMER, S. Y., DAPRETTO, M., DEEN, B., DELMONTE, S., DINSTEIN, I., ERTL-WAGNER, B., FAIR, D. A., GALLAGHER, L., KENNEDY, D. P., KEOWN, C. L., KEYSERS, C., LAINHART, J. E., LORD, C., LUNA, B., MENON, V., MINSHEW, N. J., MONK, C. S., MUELLER, S., MULLER, R.-A., NEBEL, M. B., NIGG, J. T., O’HEARN, K., PELPHREY, K. A., PELTIER, S. J., RUDIE, J. D., SUNAERT, S., THIOUX, M., TYSZKA, J. M., UDDIN, L. Q., VERHOEVEN, J. S., WENDEROTH, N., WIGGINS, J. L., MOSTOFSKY, S. H. and MILHAM, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry* **19** 659–667.
- FRIEDLAND, S. and LIM, L.-H. (2014). Computational complexity of tensor nuclear norm. *arXiv:1410.6072* .
- FRIEDMAN, J., HASTIE, H. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.
- HA, S., SOHN, I.-J., KIM, N., SIM, H. J. and CHEON, K.-A. (2015). Characteristics of brains in autism spectrum disorder: Structure, function and connectivity across the lifespan. *Experimental Neurobiology* **24** 273–284.
- HE, S., YIN, J., LI, H. and WANG, X. (2014). Graphical model selection and estimation for high dimensional tensor data. *Journal of Multivariate Analysis* **128** 165–185.
- KANA, R. K., KELLER, T. A., CHERKASSKY, V. L., MINSHEW, N. J. and JUST, M. A. (2006). Sentence comprehension in autism: thinking in pictures with decreased functional connectivity. *Brain : a journal of neurology* **129** 2484–2493.
- KARATZOGLOU, A., AMATRIAIN, X., BALTRUNAS, L. and OLIVER, N. (2010). Multi-verse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *ACM Recommender Systems*.
- KIM, J., HE, Y. and PARK, H. (2014). Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization* **58** 285–319.

- KOLDA, T. and BADER, B. (2009). Tensor decompositions and applications. *SIAM Review* **51** 455–500.
- KRUSKAL, J. B. (1976). More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* **41** 281–293.
- KRUSKAL, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications* **18** 95–138.
- LI, L. and ZHANG, X. (2016). Parsimonious tensor response regression. *Journal of the American Statistical Association* To appear.
- LI, Y., ZHU, H., SHEN, D., LIN, W., GILMORE, J. H. and IBRAHIM, J. G. (2011). Multiscale adaptive regression models for neuroimaging data. *Journal of the Royal Statistical Society: Series B* **73** 559–578.
- LIU, J., LIU, J., WONKA, P. and YE, J. (2012). Sparse non-negative tensor factorization using columnwise coordinate descent. *Pattern Recognition* **45** 649–656.
- LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35** 208–220.
- MASSART, P. (2003). Concentration inequalities and model selection. *Ecole d’Ete de Probabilites de Saint-Flour 23* .
- PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression model. *Journal of the American Statistical Association* **104** 735–746.
- RASKUTTI, G. and YUAN, M. (2016). Convex regularization for high-dimensional tensor regression. *arXiv* .
- RUDIE, J., BROWN, J., BECK-PANCER, D., HERNANDEZ, L., DENNIS, E., THOMPSON, P., BOOKHEIMER, S. and DAPRETTO, M. (2013). Altered functional and structural brain network organization in autism. *NeuroImage: Clinical* **2** 79 – 94.
- SHI, R. and KANG, J. (2015). Thresholded multiscale gaussian processes with application to bayesian feature selection for massive neuroimaging data. *arXiv preprint arXiv:1504.06074v2* .
- SIGNORETTO, M., DINH, Q., LATHAUWER, L. and SUYKENS, J. (2014). Learning with tensors: A framework based on convex optimization and spectral regularization. *Machine Learning* **93** 303–351.
- SLEPIAN, D. (1962). The one-sided barrier problem for gaussian noise. *The Bell System Technical Journal* **41** 463–501.
- SUN, W., LU, J., LIU, H. and CHENG, G. (2016). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society, Series B* To Appear.



- SUN, W., WANG, Z., LIU, H. and CHENG, G. (2015). Non-convex statistical optimization for sparse tensor graphical model. *Advances in Neural Information Processing Systems* .
- TRAVERS, B. G., KANA, R. K., KLINGER, L. G., KLEIN, C. L. and KLINGER, M. R. (2015). Motor learning in individuals with autism spectrum disorder: Activation in superior parietal lobule related to learning and repetitive behaviors. *Autism Research* **8** 38–51.
- TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in {SPM} using a macroscopic anatomical parcellation of the {MNI} {MRI} single-subject brain. *NeuroImage* **15** 273 – 289.
- WANG, X., NAN, B., ZHU, J. and KOEPPE, R. (2014a). Regularized 3D functional regression for brain image data via haar wavelets. *The Annals of Applied Statistics* **8** 1045–1064.
- WANG, X. and ZHU, H. (2016). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association* to appear.
- WANG, Y., KANG, J., KEMMER, P. B. and GUO, Y. (2016). An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Frontiers in Neuroscience* **10** 1–17.
- WANG, Z., GU, Q., NING, Y. and LIU, H. (2014b). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729* .
- YI, X. and CARAMANIS, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. *arXiv preprint* .
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.
- YUAN, M. and ZHANG, C. (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics* **16** 1031–1068.
- ZHENG, N., LI, Q., LIAO, S. and ZHANG, L. (2010). Flickr group recommendation based on tensor decomposition. In *International ACM SIGIR Conference*.
- ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* **108** 540–552.
- ZHU, H., CHEN, Y., IBRAHIM, J. G., LI, Y., HALL, C. and LIN, W. (2009). Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association* **104** 1203–1212.
- ZHU, H., KHONDKER, Z., LU, Z. and IBRAHIM, J. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association* **109** 997–990.

# Sparse Tensor Response Regression and Neuroimaging Analysis

## Supplementary Materials

Will Wei Sun <sup>1</sup> and Lexin Li <sup>2</sup>

In this online supplementary note, we provide technical proofs of all lemmas and corollaries.

### S.1 Proof of Lemma 1

To solve (4), we can use the alternating updating method to update one parameter at a time. In particular, for each  $j = 1, \dots, m$ , given  $\beta_k^{(j')}$  with  $j' \neq j$ , we solve

$$\hat{\beta}_{k,j} := \arg \min_{\substack{\beta_{k,j} \\ \|\beta_{k,j}\|_2=1, \|\beta_{k,j}\|_0 \leq s_j}} \frac{1}{n} \sum_{i=1}^n \alpha_{ik}^2 \left\| \mathcal{R}_i - w_k \beta_{k,1} \circ \dots \circ \beta_{k,m} \right\|_F^2.$$

According to the matrix representation of tensor operations (Kolda and Bader, 2009; Kim et al., 2014), this optimization problem is equivalent to solve

$$\min_{\substack{\beta_{k,j} \\ \|\beta_{k,j}\|_2=1, \|\beta_{k,j}\|_0 \leq s_j}} \frac{1}{n} \sum_{i=1}^n \alpha_{ik}^2 \left\| [\mathcal{R}_i]_{(j)}^\top - \mathbf{h}_k^{(j)} \beta_k^{(j)\top} \right\|_F^2, \quad (\text{S.1})$$

where  $[\mathcal{R}_i]_{(j)}$  is the mode- $j$  matricization of the tensor  $\mathcal{R}_i$ , and

$$\mathbf{h}_k^{(j)} := \beta_{k,m} \odot \dots \odot \beta_k^{(j+1)} \odot \beta_k^{(j-1)} \odot \dots \odot \beta_{k,1} \in \mathbb{R}^{\prod_{j' \neq j} d_{j'}}$$

with  $\odot$  the Khatri-Rao product. According to Lemma 3, we can conclude that the minimizer of (S.1) is also the solution of

$$\min_{\substack{\beta_{k,j} \\ \|\beta_{k,j}\|_2=1, \|\beta_{k,j}\|_0 \leq s_j}} \left\| \left[ \frac{1}{n} \sum_{i=1}^n \alpha_{ik}^2 \mathcal{R}_i \right]_{(j)}^\top - \mathbf{h}_k^{(j)} \beta_k^{(j)\top} \right\|_F^2.$$

This is equivalent to solve

$$\min_{\substack{\beta_{k,j} \\ \|\beta_{k,j}\|_2=1, \|\beta_{k,j}\|_0 \leq s_j}} \left\| \frac{1}{n} \sum_{i=1}^n \alpha_{ik}^2 \mathcal{R}_i - w_k \beta_{k,1} \circ \dots \circ \beta_{k,m} \right\|_F^2.$$

given other parameters  $\beta_k^{(j')}$  with  $j' \neq j$ . This leads to the desirable result in Lemma 1.  $\square$

---

<sup>1</sup>Assistant Professor, Department of Management Science, University of Miami School of Business Administration, Miami, FL 33146. Email: wsun@bus.miami.edu.

<sup>2</sup>Associate Professor, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-3370. Email: lexinli@berkeley.edu.

## S.2 Proof of Lemma 2

Denote  $[t_i]_{i_1, \dots, i_m}$  as the  $(i_1, \dots, i_m)$ -th entry of the tensor  $\mathcal{T}_i$ , and denote  $a_{i_1, \dots, i_m}$  as the  $(i_1, \dots, i_m)$ -th entry of the tensor  $\mathcal{A}_k$ . By the definition of the tensor Frobenius norm, we have

$$\frac{\partial \frac{1}{n} \sum_{i=1}^n \left\| \mathcal{T}_i - \alpha^\top \mathbf{x}_i \mathcal{A}_k \right\|_F^2}{\partial \alpha} = \frac{1}{n} \sum_{i=1}^n \sum_{i_1, \dots, i_m} 2([t_i]_{i_1, \dots, i_m} - \alpha^\top \mathbf{x}_i a_{i_1, \dots, i_m})(-a_{i_1, \dots, i_m} \mathbf{x}_i).$$

Setting the above gradient to zero implies that

$$\frac{1}{n} \sum_{i=1}^n \left\{ \sum_{i_1, \dots, i_m} [t_i]_{i_1, \dots, i_m} a_{i_1, \dots, i_m} \right\} \mathbf{x}_i = \frac{1}{n} \sum_{i=1}^n \sum_{i_1, \dots, i_m} \alpha^\top \mathbf{x}_i a_{i_1, \dots, i_m}^2 \mathbf{x}_i,$$

where the left-hand-side equals  $n^{-1} \sum_{i=1}^n \|\mathcal{T}_i * \mathcal{A}_k\|_+ \mathbf{x}_i$ , and the right-hand-side equals

$$\left\{ \sum_{i_1, \dots, i_m} a_{i_1, \dots, i_m}^2 \right\} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \alpha = \|\mathcal{A}_k\|_F^2 \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \alpha.$$

This leads to the desirable conclusion in Lemma 2.  $\square$

## S.3 Proof of Lemma 3

We first derive the solution of

$$\arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n \gamma_i^2 \left\| \mathbf{Y}_i - \alpha \beta^\top \right\|_F^2,$$

and then connect it with the optimization problem on the right-hand-side. Denote  $[\mathbf{Y}_i]_{sj} \in \mathbb{R}$  as the  $(s, j)$ -th entry of the matrix  $\mathbf{Y}_i$ , and denote  $[\mathbf{Y}_i]_j \in \mathbb{R}^{d_1}$  as the  $j$ -th column of the matrix  $\mathbf{Y}_i$ . Note that, for each  $j = 1, \dots, d_2$ , solving

$$\frac{\partial \frac{1}{n} \sum_{i=1}^n \gamma_i^2 \left\| \mathbf{Y}_i - \alpha \beta^\top \right\|_F^2}{\partial \beta_j} = \frac{1}{n} \sum_{i=1}^n \sum_{s=1}^{d_1} 2\gamma_i^2 ([\mathbf{Y}_i]_{sj} - \alpha_s \beta_j)(-\alpha_s) = 0$$

leads to the solution

$$\hat{\beta}_j = \frac{\frac{1}{n} \sum_{i=1}^n \sum_{s=1}^{d_1} \gamma_i^2 [\mathbf{Y}_i]_{sj} \alpha_s}{\frac{1}{n} \sum_{i=1}^n \gamma_i^2 \sum_{s=1}^{d_1} \alpha_s^2} = \frac{\frac{1}{n} \sum_{i=1}^n \gamma_i^2 [\mathbf{Y}_i]_j^\top \alpha}{(\frac{1}{n} \sum_{i=1}^n \gamma_i^2) \alpha^\top \alpha}.$$

The solution  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_{d_2})^\top$  is indeed a minimizer to the original optimization problem by noting that the second order derivative of  $\beta_j$  is positive. By solving a similar equation, we get the minimizer of the optimization problem  $\min_{\beta_j} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i - \alpha \beta^\top \right\|_F^2$  as

$$\tilde{\beta}_j = \frac{\frac{1}{n} \sum_{i=1}^n \gamma_i^2 [\mathbf{Y}_i]_j^\top \alpha}{\alpha^\top \alpha}.$$

Note that  $\widehat{\beta}_j$  equals to  $\widetilde{\beta}_j$  up to a constant  $n^{-1} \sum_{i=1}^n \gamma_i^2$ . Therefore, normalizing  $\widehat{\beta}_j$  and  $\widetilde{\beta}_j$  to be unit-norm vector leads to the same solution. This completes the proof of Lemma 3.  $\square$

## S.4 Proof of Lemma 9

The key step is to verify the conditions in Lemma 8. In our example, the incoherence condition, Assumption 4, trivially holds, since the incoherence parameter  $\zeta = 0$ . Therefore, the function  $f(\epsilon; K, d_0) = C_3 \epsilon^2$  in our example. According to Lemma 8, we have

$$D(\widehat{\beta}_{k,3}, \beta_{k,3}^*) \leq \frac{\sqrt{5} w_{\max}^* C_3 \epsilon^2 + \sqrt{5} \eta(\widehat{\mathcal{E}}, d_0 + s)}{w_j(1 - \epsilon^2) - w_{\max}^* C_3 \epsilon^2 - \eta(\widehat{\mathcal{E}}, d_0 + s)}.$$

We next simplify the denominator  $w_j(1 - \epsilon^2) - w_{\max}^* C_3 \epsilon^2 - \eta(\widehat{\mathcal{E}}, d_0 + s)$ . By the assumption on  $\epsilon$ , we have

$$w_j(1 - \epsilon^2) - w_{\max}^* C_3 \epsilon^2 \geq w_{\min}^*/2,$$

which, together with the condition on the error tensor  $\eta(\widehat{\mathcal{E}}, d_0 + s)$ , implies that

$$w_j(1 - \epsilon^2) - w_{\max}^* C_3 \epsilon^2 - \eta(\widehat{\mathcal{E}}, d_0 + s) \geq w_{\min}^*/4.$$

Therefore, we have

$$D(\widehat{\beta}_{k,3}, \beta_{k,3}^*) \leq \frac{4\sqrt{5} w_{\max}^* C_3}{w_{\min}^*} \epsilon^2 + \frac{4\sqrt{5}}{w_{\min}^*} \eta(\widehat{\mathcal{E}}, d_0 + s),$$

which completes the proof of Lemma 9.  $\square$

## S.5 Proof of Lemma 10

According to the explicit form in (14), we have

$$\eta(\widehat{\mathcal{E}}, s) \leq \eta(I_1, s) + \eta(I_2, s).$$

Next we bound the two terms  $\eta(I_1, s)$  and  $\eta(I_2, s)$  separately.

**Bound  $\eta(I_1, s)$ :** Denote  $\mathbb{S}_q^d = \{\mathbf{v} \in \mathbb{R}^d \mid \|\mathbf{v}\|_2 = 1, \|\mathbf{v}\|_0 \leq q\}$ . According to the definiteness of  $\eta(\cdot)$ , we have

$$\begin{aligned} \eta(I_1, s) &= \sup_{\mathbf{u} \in \mathbb{S}_s^{d_1}, \mathbf{v} \in \mathbb{S}_s^{d_2}, \mathbf{w} \in \mathbb{S}_s^{d_3}} |I_1 \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w}| \\ &\leq \max_i \left| \frac{(\beta_{k,4}^* - \widehat{\beta}_{k,4})^\top \mathbf{x}_i}{\widehat{\beta}_{k,4}^\top \mathbf{x}_i \beta_{k,4}^{*\top} \mathbf{x}_i} \right| \sup_{\mathbf{u} \in \mathbb{S}_s^{d_1}, \mathbf{v} \in \mathbb{S}_s^{d_2}, \mathbf{w} \in \mathbb{S}_s^{d_3}} \left| \left( \frac{1}{n} \sum_{i=1}^n \mathcal{Y}_i \right) \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} \right|. \end{aligned}$$

By the Cauchy Schwarz inequality, we have  $\|(\beta_{k,4}^* - \widehat{\beta}_{k,4})^\top \mathbf{x}_i\| \leq \|\beta_{k,4}^* - \widehat{\beta}_{k,4}\| \|\mathbf{x}_i\| \leq C_2 \|\beta_{k,4}^* - \widehat{\beta}_{k,4}\| \leq C_2 \epsilon$ . Moreover, note that  $|\widehat{\beta}_{k,4}^\top \mathbf{x}_i \beta_{k,4}^{*\top} \mathbf{x}_i| = |\beta_{k,4}^{*\top} \mathbf{x}_i + (\widehat{\beta}_{k,4} - \beta_{k,4}^*)^\top \mathbf{x}_i| |\beta_{k,4}^{*\top} \mathbf{x}_i|$ .

When the element of  $\mathbf{x}_i$  is bounded for each  $i = 1, \dots, n$  and the dimension  $d_{m+1}$  of  $\mathbf{x}_i$  is fixed, we have  $\|\mathbf{x}_i\| \leq C_2$ . Then by the assumption that  $|\boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i| \geq C_3$  and  $\|\mathbf{x}_i\| \leq C_2$ , we have

$$|\widehat{\boldsymbol{\beta}}_{k,4}^\top \mathbf{x}_i \boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i| \geq (C_3 - C_2 \|\widehat{\boldsymbol{\beta}}_{k,4} - \boldsymbol{\beta}_{k,4}^*\|) C_3 \geq (C_3 - C_2 \epsilon) C_3 \geq C_3^2/2, \quad (\text{S.2})$$

where the last inequality is due to the assumption that  $\epsilon \leq C_3/(2C_2)$ . Then,

$$\max_i \left| \frac{(\boldsymbol{\beta}_{k,4}^* - \widehat{\boldsymbol{\beta}}_{k,4})^\top \mathbf{x}_i}{\widehat{\boldsymbol{\beta}}_{k,4}^\top \mathbf{x}_i \boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i} \right| \leq \frac{2C_2\epsilon}{C_3^2}.$$

Moreover, according to the true model  $\mathcal{Y}_i = w^*(\boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i) \boldsymbol{\beta}_{k,1}^* \circ \boldsymbol{\beta}_{k,2}^* \circ \boldsymbol{\beta}_{k,3}^* + \mathcal{E}_i$ ,

$$\begin{aligned} & \sup_{\mathbf{u} \in \mathbb{S}_s^{d_1}, \mathbf{v} \in \mathbb{S}_s^{d_2}, \mathbf{w} \in \mathbb{S}_s^{d_3}} \left| \left( \frac{1}{n} \sum_{i=1}^n \mathcal{Y}_i \right) \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} \right| \\ & \leq \sup_{\mathbf{u} \in \mathbb{S}_s^{d_1}, \mathbf{v} \in \mathbb{S}_s^{d_2}, \mathbf{w} \in \mathbb{S}_s^{d_3}} \left| \left( w^* \left( \boldsymbol{\beta}_{k,4}^{*\top} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right) \boldsymbol{\beta}_{k,1}^* \circ \boldsymbol{\beta}_{k,2}^* \circ \boldsymbol{\beta}_{k,3}^* \right) \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} \right| + \eta(\bar{\mathcal{E}}, s) \\ & = \eta(\bar{\mathcal{E}}, s), \end{aligned}$$

where the last equality is due to the fact that the covariate vector  $\mathbf{x}_i$  is centered such that  $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$ . Therefore, we have,

$$\eta(I_1, s) \leq \underbrace{\frac{2C_2\eta(\bar{\mathcal{E}}, s)}{C_3^2}}_{\kappa_2} \epsilon.$$

**Bound  $\eta(I_2, s)$ :** By the assumption that, for each  $i$ ,  $|\boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i| \geq C_3$ , we have that  $\max_i (|\boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i|)^{-1} \leq 1/C_3$ , and hence

$$\begin{aligned} \eta(I_2, s) &= \sup_{\mathbf{u} \in \mathbb{S}_s^{d_1}, \mathbf{v} \in \mathbb{S}_s^{d_2}, \mathbf{w} \in \mathbb{S}_s^{d_3}} \left| \left( \frac{1}{n} \sum_{i=1}^n \frac{\mathcal{E}_i}{\boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i} \right) \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} \right| \\ &\leq \max_i (|\boldsymbol{\beta}_{k,4}^{*\top} \mathbf{x}_i|)^{-1} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) \\ &\leq \frac{1}{C_3} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right). \end{aligned}$$

Combining the above two results, we obtain the desirable upper bound of  $\eta(\widehat{\mathcal{E}}, s)$ .  $\square$

## S.6 Proof of Lemma 11

**Bound (I):** For any two tensors  $\mathcal{A}, \mathcal{B}$  of the same dimension, the Cauchy-Schwarz inequality implies that  $\langle \mathcal{A}, \mathcal{B} \rangle \leq \|\mathcal{A}\|_F \|\mathcal{B}\|_F$ . Therefore, we have

$$(I) \leq \frac{\|\hat{\mathcal{A}}_k - \mathcal{A}_k^*\|_F}{\|\hat{\mathcal{A}}_k\|_F}.$$

We next simplify the above numerator  $\|\hat{\mathcal{A}}_k - \mathcal{A}_k^*\|_F$ . By definition,

$$\begin{aligned} \|\hat{\mathcal{A}}_k - \mathcal{A}_k^*\|_F &= \left\| \hat{w}_k \hat{\beta}_{k,1} \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} - w_k^* \beta_{k,1}^* \circ \beta_{k,2}^* \circ \beta_{k,3}^* \right\|_F \\ &\leq \underbrace{\left\| \hat{w}_k \hat{\beta}_{k,1} \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} - w_k^* \hat{\beta}_{k,1} \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} \right\|_F}_{I_1} \\ &\quad + \underbrace{\left\| w_k^* \hat{\beta}_{k,1} \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} - w_k^* \beta_{k,1}^* \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} \right\|_F}_{I_2} \\ &\quad + \underbrace{\left\| w_k^* \beta_{k,1}^* \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} - w_k^* \beta_{k,1}^* \circ \beta_{k,2}^* \circ \hat{\beta}_{k,3} \right\|_F}_{I_3} \\ &\quad + \underbrace{\left\| w_k^* \beta_{k,1}^* \circ \beta_{k,2}^* \circ \hat{\beta}_{k,3} - w_k^* \beta_{k,1}^* \circ \beta_{k,2}^* \circ \beta_{k,3}^* \right\|_F}_{I_4}. \end{aligned}$$

We next bound each term  $I_j / \|\hat{\mathcal{A}}_k\|_F$  for  $j = 1, 2, 3, 4$ . First, according to the assumptions  $|\hat{w}_k - w_k^*| \leq \epsilon$  and  $\epsilon \leq w_{\min}^*/2$ , we have  $|\hat{w}_k| \geq w_k^* - \epsilon \geq w_{\min}^*/2$ . Therefore, we have

$$\frac{I_1}{\|\hat{\mathcal{A}}_k\|_F} = \left| \frac{\hat{w}_k - w_k^*}{\hat{w}_k} \right| \leq \frac{\epsilon}{|\hat{w}_k|} \leq \frac{2}{w_{\min}^*} \epsilon.$$

In addition, note that for any vectors  $\mathbf{a} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^n$ , since the rank of the matrix  $\mathbf{a} \circ \mathbf{b}$  is 1, we have  $\|\mathbf{a} \circ \mathbf{b}\|_F = \|\mathbf{a} \circ \mathbf{b}\|_2$ . This equality and the fact that  $\|\beta_{k,j}^*\|_2 = \|\hat{\beta}_{k,j}\|_2 = 1$  for any  $j = 1, 2, 3$  imply that

$$\frac{I_2}{\|\hat{\mathcal{A}}_k\|_F} = \frac{\left\| w_k^* (\hat{\beta}_{k,1} - \beta_{k,1}^*) \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} \right\|_F}{\left\| \hat{w}_k \hat{\beta}_{k,1} \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3} \right\|_F} \leq \frac{|w_k^*| \|\hat{\beta}_{k,1} - \beta_{k,1}^*\|_2}{|\hat{w}_k|} \leq 2\sqrt{2}\epsilon,$$

where the last inequality is due to the inequality  $|\hat{w}_k| \geq w_{\min}^*/2$  as we show above, as well as the fact that  $D(\mathbf{u}, \mathbf{v}) \leq \min\{\|\mathbf{u} - \mathbf{v}\|_2, \|\mathbf{u} + \mathbf{v}\|_2\} \leq \sqrt{2}D(\mathbf{u}, \mathbf{v})$  for unit vectors  $\mathbf{u}, \mathbf{v}$ . By applying similar proof techniques, we also have

$$\frac{I_3}{\|\hat{\mathcal{A}}_k\|_F} \leq 2\sqrt{2}\epsilon; \quad \frac{I_4}{\|\hat{\mathcal{A}}_k\|_F} \leq 2\sqrt{2}\epsilon.$$

Combing the above four inequalities, we obtain the desirable bound for (I) in (23).

**Bound (II):** It is easy to see that

$$(II) \leq \frac{\frac{1}{n} \sum_{i=1}^n \langle \mathcal{E}_i, \hat{\mathcal{A}}_k \rangle \|\Omega \mathbf{x}_i\|_2}{\|\hat{\mathcal{A}}_k\|_F^2}.$$

According to the assumptions that  $\Omega$  is positive definite with bounded eigenvalues and  $\|\mathbf{x}_i\| \leq C_2$  for each  $i$ , we have  $\max_i \|\Omega \mathbf{x}_i\|_2$  is bounded by some constant  $C_1 > 0$ . In addition, according to the Gaussian comparison inequality Lemma 4, as well as the large deviation bound, we have that, with high probability

$$\frac{1}{n} \sum_{i=1}^n \langle \mathcal{E}_i, \hat{\mathcal{A}}_k \rangle \leq \mathbb{E} \left[ \frac{1}{\sqrt{n}} \langle \mathcal{G}, \hat{\mathcal{A}}_k \rangle \right] \leq \frac{1}{\sqrt{n}} \mathbb{E} \left[ \|\mathcal{G}\|_F \|\hat{\mathcal{A}}_k\|_F \right],$$

for some Gaussian tensor  $\mathcal{G} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$  whose entries are i.i.d. standard normal random variables. Therefore, we have, with high probability, the desirable inequality in (24) holds, where the whole term  $\tilde{C}$  is bounded by noting that  $\max_i \|\Omega \mathbf{x}_i\|_2$  is bounded and all entries of tensors  $\mathcal{G}$ , and  $\hat{\mathcal{A}}_k$  of the same dimension are bounded.  $\square$

## S.7 Proof of Corollary 1

The derivation of  $\eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right)$  in the Gaussian error tensor scenario consists of three parts. In Stage 1, we apply Lemma 4 to show that

$$\mathbb{P} \left[ \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) > x \right] \leq \mathbb{P} \left[ n^{-1/2} \eta(\mathcal{G}, s) > x \right],$$

for some Gaussian tensor  $\mathcal{G}$ . In Stage 2, we show via the large deviation bound inequality that

$$\mathbb{P} (|\eta(\mathcal{G}, s) - \mathbb{E}[\eta(\mathcal{G}, s)]| \geq t) \leq 2 \exp \left( -\frac{t^2}{2L^2} \right)$$

with some Lipschitz constant  $L$ . In Stage 3, by incorporating Lemma 6, and exploring the sparsity constraint, we can compute

$$\mathbb{E}[\eta(\mathcal{G}, s)] \leq C \sqrt{s^3 \log(d_1 d_2 d_3)},$$

for some constant  $C > 0$ .

**Stage 1:** According to the definition of  $\eta(\cdot)$  in (8), we have

$$\begin{aligned} \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) &= \sup_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1 \\ \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s, \|\mathbf{w}\|_0 \leq s}} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} \right| \\ &= \sup_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1 \\ \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s, \|\mathbf{w}\|_0 \leq s}} \left\langle \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \right\rangle. \end{aligned}$$

Clearly, for any unit-norm vectors  $\mathbf{u}, \mathbf{v}, \mathbf{w}$ , when  $\mathcal{E}_i, i = 1, \dots, n$  are i.i.d. Gaussian tensors, we have

$$\mathbb{E} \left[ \left\langle \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \right\rangle \right] = 0.$$

Since the variance of each entry of  $\mathcal{E}_i$  is 1, we also have

$$\begin{aligned} \text{var} \left[ \left\langle \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \right\rangle \right] &= \text{var} \left[ \frac{1}{n} \sum_{i=1}^n \langle \mathcal{E}_i, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \rangle \right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{r,s,t} [\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}]_{r,s,t}^2 \text{var}([\mathcal{E}_i]_{r,s,t}) = \frac{\|\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}\|_F^2}{n}. \end{aligned}$$

For some Gaussian tensor  $\mathcal{G}$  of the same dimension as  $\mathcal{E}_i$ , it is easy to show that

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{\sqrt{n}} \langle \mathcal{G}, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \rangle \right] &= 0 \\ \text{var} \left[ \frac{1}{\sqrt{n}} \langle \mathcal{G}, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \rangle \right] &= \frac{\|\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}\|_F^2}{n}. \end{aligned}$$

Moreover, for any  $\mathcal{A}_1 := \mathbf{u}_1 \circ \mathbf{v}_1 \circ \mathbf{w}_1$  and  $\mathcal{A}_2 := \mathbf{u}_2 \circ \mathbf{v}_2 \circ \mathbf{w}_2$ , we have

$$\text{var} \left[ \left\langle \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, \mathcal{A}_1 - \mathcal{A}_2 \right\rangle \right] = \frac{\|\mathcal{A}_1 - \mathcal{A}_2\|_F}{n} = \text{var} \left[ \frac{1}{\sqrt{n}} \langle \mathcal{G}, \mathcal{A}_1 - \mathcal{A}_2 \rangle \right].$$

Therefore, according to Lemma 4, we have, for each  $x > 0$ ,

$$\mathbb{P} \left[ \eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) > x \right] \leq \mathbb{P} \left[ \frac{1}{\sqrt{n}} \eta(\mathcal{G}, s) > x \right]. \quad (\text{S.3})$$

**Stage 2:** We show that the function  $\eta(\cdot, s)$  is a Lipschitz function in its first argument. For any two tensors  $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ , denote  $\mathcal{A}^* = \sup_{\mathcal{A}} \langle \mathcal{G}_1, \mathcal{A} \rangle$ . We have

$$\sup_{\mathcal{A}} \langle \mathcal{G}_1, \mathcal{A} \rangle - \sup_{\mathcal{A}} \langle \mathcal{G}_2, \mathcal{A} \rangle \leq \langle \mathcal{G}_1, \mathcal{A}^* \rangle - \sup_{\mathcal{A}} \langle \mathcal{G}_2, \mathcal{A} \rangle \leq \langle \mathcal{G}_1, \mathcal{A}^* \rangle - \langle \mathcal{G}_2, \mathcal{A}^* \rangle \leq \langle \mathcal{G}_1 - \mathcal{G}_2, \mathcal{A}^* \rangle.$$

Therefore, we have

$$\begin{aligned} &|\eta(\mathcal{G}_1, s) - \eta(\mathcal{G}_2, s)| \\ &= \sup_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1 \\ \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s, \|\mathbf{w}\|_0 \leq s}} \left| \mathcal{G}_1 \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} \right| - \sup_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1 \\ \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s, \|\mathbf{w}\|_0 \leq s}} \left| \mathcal{G}_2 \times_1 \mathbf{u} \times_2 \mathbf{v} \times_3 \mathbf{w} \right| \\ &\leq \sup_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1 \\ \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s, \|\mathbf{w}\|_0 \leq s}} \left\langle \mathcal{G}_1 - \mathcal{G}_2, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \right\rangle \\ &\leq \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1} \|\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}\|_F \cdot \|\mathcal{G}_1 - \mathcal{G}_2\|_F \\ &\leq \|\mathcal{G}_1 - \mathcal{G}_2\|_F, \end{aligned}$$



where the second inequality is due to the fact that  $\langle \mathcal{A}, \mathcal{B} \rangle \leq \|\mathcal{A}\|_F \|\mathcal{B}\|_F$ , and the third inequality is because  $\|\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}\|_F = \|\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}\|_2 \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \|\mathbf{w}\|_2 = 1$  for any unit-norm vectors  $\mathbf{u}, \mathbf{v}, \mathbf{w}$ .

Applying the concentration result of Lipschitz functions of Gaussian random variables in Lemma 5 with  $L = 1$ , for the Gaussian tensor  $\mathcal{G}$ , we have

$$\mathbb{P}(|\eta(\mathcal{G}, s) - \mathbb{E}[\eta(\mathcal{G}, s)]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2}\right). \quad (\text{S.4})$$

**Stage 3:** We aim to bound  $\mathbb{E}[\eta(\mathcal{G}, s)]$ . For a tensor  $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ , denote its  $L_1$ -norm regularizer  $R(\mathcal{T}) = \sum_{j_1} \sum_{j_2} \sum_{j_3} |\mathcal{T}_{j_1, j_2, j_3}|$ . Define the ball of this regularizer as  $B_R(\delta) := \{\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3} | R(\mathcal{T}) \leq \delta\}$ .

For any vectors  $\mathbf{u} \in \mathbb{R}^{d_1}, \mathbf{v} \in \mathbb{R}^{d_2}, \mathbf{w} \in \mathbb{R}^{d_3}$  satisfying  $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1, \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s$ , and  $\|\mathbf{w}\|_0 \leq s$ , denote  $\mathcal{A} := \mathbf{u} \circ \mathbf{v} \circ \mathbf{w}$ . Lemma 7 implies that  $R(\mathcal{A}) \leq s^{3/2}$ .

Therefore, we have

$$\begin{aligned} \mathbb{E}[\eta(\mathcal{G}, s)] &= \mathbb{E} \left[ \sup_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=\|\mathbf{w}\|=1 \\ \|\mathbf{u}\|_0 \leq s, \|\mathbf{v}\|_0 \leq s, \|\mathbf{w}\|_0 \leq s}} \langle \mathcal{G}, \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \rangle \right] \\ &\leq \mathbb{E} \left[ \sup_{\mathcal{A} \in B_R(s^{3/2})} \langle \mathcal{G}, \mathcal{A} \rangle \right] = s^{3/2} \mathbb{E} \left[ \sup_{\mathcal{A} \in B_R(1)} \langle \mathcal{G}, \mathcal{A} \rangle \right]. \end{aligned}$$

This result, together with Lemma 6, implies that

$$\mathbb{E}[\eta(\mathcal{G}, s)] \leq C \sqrt{s^3 \log(d_1 d_2 d_3)}. \quad (\text{S.5})$$

Finally, combining (S.4) and (S.5), and setting  $t = \sqrt{s^3 \log(d_1 d_2 d_3)}$ , we have, with probability  $1 - 2 \exp(-s^3 \log(d_1 d_2 d_3)/2)$ ,

$$|\eta(\mathcal{G}, s) - \mathbb{E}[\eta(\mathcal{G}, s)]| \leq \sqrt{s^3 \log(d_1 d_2 d_3)}.$$

Henceforth

$$\eta(\mathcal{G}, s) \leq (C + 1) \sqrt{s^3 \log(d_1 d_2 d_3)},$$

which together with (S.3), implies that

$$\eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i, s \right) = O_p \left( \sqrt{\frac{s^3 \log(d_1 d_2 d_3)}{n}} \right).$$

This completes the proof of Corollary 1. □

## S.8 Proof of Corollary 2

The proof follows by incorporating the result in Lemma 1 and the structure assumption in (11). In particular, Lemma 1 implies that, when  $\tilde{\mathcal{E}}_i \in \mathbb{R}^{d \times d}$  is a matrix whose entries are i.i.d. standard Gaussian,

$$\eta \left( \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{E}}_{i,s} \right) = O_p \left( \sqrt{\frac{s^2 \log(d^2)}{n}} \right).$$

Moreover, under the structure assumption in (11), we have

$$\eta \left( \frac{1}{n} \sum_{i=1}^n \mathcal{E}_{i,s} \right) = \eta \left( \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{E}}_{i,s} \right),$$

which completes the proof of Corollary 2. □